Breaking point: the genesis and impact of structural variation in tumours

Citation for published version:
Ewing, A & Semple, C 2018, 'Breaking point: the genesis and impact of structural variation in tumours' F1000Research. DOI: 10.12688/f1000research.16079.1

Digital Object Identifier (DOI):
10.12688/f1000research.16079.1

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
F1000Research

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
BREAKING POINT: THE GENESIS AND IMPACT OF STRUCTURAL VARIATION IN TUMOURS [VERSION 1; REFEREES: 2 APPROVED]

AILITH EWING, COLIN SEMPLE

MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, The University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH42XU, UK

Abstract

Somatic structural variants undoubtedly play important roles in driving tumourigenesis. This is evident despite the substantial technical challenges that remain in accurately detecting structural variants and their breakpoints in tumours and in spite of our incomplete understanding of the impact of structural variants on cellular function. Developments in these areas of research contribute to the ongoing discovery of structural variation with a clear impact on the evolution of the tumour and on the clinical importance to the patient. Recent large whole genome sequencing studies have reinforced our impression of each tumour as a unique combination of mutations but paradoxically have also discovered similar genome-wide patterns of single-nucleotide and structural variation between tumours. Statistical methods have been developed to deconvolute mutation patterns, or signatures, that recur across samples, providing information about the mutagens and repair processes that may be active in a given tumour. These signatures can guide treatment by, for example, highlighting vulnerabilities in a particular tumour to a particular chemotherapy. Thus, although the complete reconstruction of the full evolutionary trajectory of a tumour genome remains currently out of reach, valuable data are already emerging to improve the treatment of cancer.

Keywords

Structural variation, cancer, tumourigenesis, DNA double-strand breaks, mutational signatures, whole genome sequencing
Corresponding author: Ailith Ewing (ailith.ewing@igmm.ed.ac.uk)

Author roles: Ewing A: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; Semple C: Conceptualization, Funding Acquisition, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The authors are supported by core funding of the UK Medical Research Council (MRC) to the MRC Human Genetics Unit and a UKRI Innovation Fellowship to Ailith Ewing.

Copyright: © 2018 Ewing A and Semple C. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Ewing A and Semple C. Breaking point: the genesis and impact of structural variation in tumours [version 1; referees: 2 approved] F1000Research 2018, 7(F1000 Faculty Rev):1814 (https://doi.org/10.12688/f1000research.16079.1)

First published: 19 Nov 2018, 7(F1000 Faculty Rev):1814 (https://doi.org/10.12688/f1000research.16079.1)
**Introduction**

Tumours arise when normal cells accumulate enough genetic alterations to affect normal regulatory control systems. These alterations usually include somatic structural variants (SVs) which are large-scale changes to the genome, ranging from 50 nucleotides to many megabases in length. SVs can occur as a result of improper DNA double-strand break (DSB) repair and are traditionally classified into five different types: insertions, deletions, duplications, inversions, and translocations. Although all tumours accumulate both somatic single-nucleotide variant (SNV) mutations and SVs, certain tumour types such as high-grade serous ovarian cancer (HGSOC) and invasive breast cancer are dominated by SVs. Furthermore, in HGSOC and pancreatic neuroendocrine tumours, there is good evidence for SVs playing prominent roles in driving tumourigenesis. Various sequencing strategies, particularly whole genome sequencing (WGS), can be used to identify structural variation in tumours by comparing the pattern of structural variation in tumour tissue with that observed in normal cells. A range of computational tools are then used to detect these variants from sequencing data, but detection algorithms perform very differently in identifying particular SV classes, and none of the existing tools are effective in detecting all SVs on their own. Copy number alteration (CNA) detection in cancer samples has been reviewed extensively because of its important role in tumourigenesis. For example, it has been suggested that HLA allele loss in 40% of non-small cell lung cancers enables immune escape. The detection of CNA is a more tractable problem than the detection of many other SVs, as CNAs can be inferred purely on the basis of changes in read depth over a region. Regions with lower read depth suggest loss of genetic material, whereas regions of higher read depth suggest gain. Aside from detecting copy number changes, analyses of SVs in the past have focused on investigating simpler rearrangements of known cancer genes or identifying fusion genes; however, the focus is now shifting to considering more complex events and SVs as part of broader, genome-wide patterns of mutation. Our present knowledge of structural aberrations is clearly far from complete, such that although aneuploidy is well studied, large-scale phenomena such as whole genome doubling have only recently been discovered to be a common event in many cancers and to have important implications for clinical prognoses.

Even when an SV is associated with patient survival, it is challenging to uncover the underlying mechanism (that is, how the SV impacts tumour function). It is known that SVs can alter the expression of oncogenes or tumour suppressor genes and that these changes in expression may directly contribute to tumourigenesis. It also seems that SVs with breakpoints in the main body of a gene may often be associated with increased expression of the gene but that SVs with breakpoints within 50 kb upstream or downstream of a gene are correlated with decreased expression. Similarly, it has been observed that amplified regions of the genome are enriched for oncogenes but that deleted regions are enriched for tumour suppressors. However, SVs may alter the expression of genes through various mechanisms, and multiple SVs may occur at the same locus, so the impact of a particular SV on nearby genes is often difficult to predict. SVs have long been known to result in the formation of functional fusion genes by translocations. For example, the BCR-ABL fusion gene is formed by a t(9;22)(q34;q11) translocation which leads to increased tyrosine kinase activity, uncontrolled proliferation, and ultimately chronic myeloid leukaemia (CML). This type of gene fusion can be targeted with kinase inhibitors, resulting in an effective treatment for CML. Rearrangements can also reposition regulatory elements such as enhancers next to oncogenes, which has been observed in Burkitt’s lymphoma and results in increased expression of the MYC oncogene. Recent studies have suggested that enhancer adoption by oncogenes following the disruption of regulatory domains is fairly common in tumourigenesis, occurring at rates comparable to those of recurrent in-frame gene fusions, although the domain architectures themselves may simply be prone to higher rates of mutation. Finally, there appear to be SVs with breakpoints that do not have direct effects on either genes or regulatory elements and yet their presence is associated with patient survival rates. Beyond gene expression, the effects of SVs are not well studied, but notably modest correlations have been reported between the extent of tumour duplications and deletions involving a gene and the levels of the protein it encodes. In summary, although simpler SVs may tend to have intuitive effects on gene expression (for example, deletions leading to lower expression and amplifications leading to higher expression), it is generally not straightforward to interpret the functional effects of a given SV within a given tumour.

**Identifying structural variation**

Historically, chromosomal aberrations were detected by using microscopy and traditional karyotyping methods, which were then improved with the development of fluorescent in situ hybridisation techniques. Modern high-throughput sequencing technologies are capable of detecting a wider variety of structural variation at much higher resolution. The most widely used sequencing technologies generate billions of short reads about 100 nucleotides in length from sequencing libraries. The short lengths of the reads, sequencing platform error rates, and the sequencing libraries used can all present challenges for SV detection. Some reads map ambiguously to many different regions in the reference genome, which makes it difficult to determine the true location of the corresponding SV. This is particularly true of reads in repetitive regions. Smaller library insert sizes enable the precise identification of small SVs such as insertions and deletions; however, smaller insert sizes will cover only short stretches of sequence, making it hard to cover large SVs. Contrastingly, larger insert sizes allow the identification of large SVs but may have insufficient coverage to determine the location of breakpoints at base-pair resolution. Given the high cost of conducting sequencing experiments using multiple library sizes, it is important to use the appropriate strategy for the biological hypothesis under study. Chimeric reads are artificial sequences formed during the amplification step of polymerase chain reaction (PCR) and can also result in artefacts during analysis. The effect of chimeric reads on SV prediction can be mitigated by using multiple libraries per sample and filtering out novel sequences that are in only one library; however, this comes with the risk of filtering out low-frequency SVs.
Alternatively, there are tools such as ChimeraChecker\textsuperscript{32} and Chimera Slayer\textsuperscript{33} which are designed to remove chimeric reads. Finally, the genomic context of the sequence also impacts SV calling. Regions that are poorly characterised in the reference genome such as pericentromeric and telomeric regions present a challenge to the identification of SVs. In addition, repetitive or GC-rich regions are often problematic for confident read mapping, which can be overcome by sequencing technologies generating longer reads\textsuperscript{3}.

Several long-range technologies have been developed to try to address these challenges\textsuperscript{34}. PacBio\textsuperscript{35} and Oxford Nanopore\textsuperscript{36} sequencing platforms both produce long reads that can be used for SV detection. The resultant reads from these technologies average about 10 kb in length and some read lengths recorded up to 100 kb for PacBio and 1 Mb for Oxford Nanopore. However, these technologies are more expensive than standard Illumina short-read sequencing and have higher error rates. An alternative approach is to use a technology based on linked short reads, such as the platforms developed by 10X Genomics\textsuperscript{37}. This enables the identification of large SVs at a lower cost than for the long-read technologies; however, sequencing coverage can be relatively sparse, limiting the resolution of locally repetitive sequences\textsuperscript{38}. Although they are established, these technologies are currently not widely used in cancer genomics and the majority of SV detection in tumours for the moment is still carried out on the basis of short-read sequencing data.

Most SV detection algorithms or “callers” identify SVs from paired-end sequencing reads that are anomalously mapped. Anomalous reads can be used to predict candidate SV loci because reads that do not contain an SV breakpoint are expected to map concordantly\textsuperscript{2}. Anomalously mapped reads fall into one of three categories: discordant reads, soft-clipped reads, and one-end-anchored reads. Discordant reads can be used in the direct discovery of SVs and are mapped to either different chromosomes (providing evidence for translocations) or incorrect strands (suggesting inversions) or are in incorrect orientation (suggesting duplications) or are an incorrect distance apart (indicating the presence of an insertion or deletion). Detecting SVs from discordant reads is more effective for the detection of larger SV events, as they are unable to estimate breakpoint locations precisely and small SVs may have insufficient discordant read coverage to be detectable\textsuperscript{4}. Soft-clipped and one-end-anchored reads can be used for indirect SV discovery. Soft-clipped reads are only partially mapped, and one-end-anchored reads have only one read of the pair mapped. In both cases, split-read mapping is needed to successfully map both ends of the SV. These categories of anomaly are better for detecting smaller events such as small insertions and deletions, as they are able to estimate the breakpoints at a higher resolution.

In practice, SV callers use one or more types of anomalous reads to predict candidate SV loci in the following ways\textsuperscript{2}: first, by clustering discordant reads that support the same type of SV together (for example, SVDetect\textsuperscript{3}); second, by using split-read mapping to map the other end of indirectly discovered SVs (for example, BreakSeq\textsuperscript{3}) or to refine the estimate of the breakpoint location in directly discovered SVs (for example, LUMPY\textsuperscript{9}); and, third, by using de novo contig assembly to stitch together short reads to make a longer segment that can be mapped back to the reference and improve the ability to locate breakpoints (for example, CREST\textsuperscript{39}). Some callers use combinations of all of these techniques, in addition to assessing changes in read depth, in order to improve their detection sensitivity (for example, DELLY\textsuperscript{42}).

Assessing the accuracy of SV callers has proven to be challenging. The International Cancer Genome Consortium–The Cancer Genome Atlas (ICG-TCGA) DREAM SMC-DNA Challenge (https://www.synapse.org/#!Synapse:syn312572/wiki/) was launched in 2013, aiming to benchmark somatic SNV and SV predictions from a wide variety of algorithms using simulated and real tumour-normal WGS pairs. The project revealed widespread biases in caller predictions and recommended the use of consensus calls made by multiple callers for both SNV and SV prediction\textsuperscript{41}. More detailed accounts of the SV callers’ performance showed that several algorithms (including Breakdancer, DELLY, Findel, Manta, and novoBreak) could achieve high precision after stringent filtering of SV calls and substantial overlaps in the breakpoints called between them\textsuperscript{42}. Unfortunately, SV calling is a rapidly evolving field, and there is no ongoing attempt to comprehensively benchmark algorithms as they appear or are updated. In spite of this, multiple comparative reviews of the most popular SV detection tools and algorithms have emerged\textsuperscript{43,44}. The ensemble approach to both SNV and SV calling explored in the DREAM Challenge has since become the de facto standard operating procedure in the field, as reflected in the recent re-analyses of large TCGA\textsuperscript{45} and ICGC\textsuperscript{46} tumour sequencing studies. The choice of algorithms employed in such an ensemble is often influenced by practical considerations, such as computational demands and run times, but should also consider the shared assumptions and biases between tools\textsuperscript{47}. Because any ensemble is expected to generate false-positive and false-negative results, there is also a requirement for validation by orthogonal data for SVs of particular interest, and long-read sequencing technologies may often be useful in this respect\textsuperscript{48}.

**Genesis and patterns of structural variation**

SVs arise when DNA DSBs are improperly repaired\textsuperscript{51}. The three main DSB repair mechanisms are homologous recombination (HR), non-homologous end joining (NHEJ), and replication-based mechanisms\textsuperscript{52}. HR is the most common DNA repair mechanism\textsuperscript{1} and is accurate as long as there is a homologous sequence to use as a repair template. However, in tumour genomes, there are few overlapping sequences at the breakpoints\textsuperscript{53,54}, so generally HR is not possible and the error-prone NHEJ is implemented instead. NHEJ requires no homology and sometimes can generate short segments of microhomology or small insertions at the breakpoints\textsuperscript{5}. Replication-based mechanisms of DNA DSB repair such as microhomology-mediated break-induced repair (MBMIR) have also been described\textsuperscript{55}. It is thought that in these cases the replication fork can stall and the polymerase can switch template sequence using microhomology to any nearby single-stranded DNA\textsuperscript{5}. This can result in a range of complex structural rearrangements\textsuperscript{5,6,7}. In tumour genomes, which are often HR
deficient, these last two mechanisms of repair dominate. A further mechanism behind structural variation in cancer is chromosome instability. This is defined as an increased rate of chromosomal change in comparison with normal cells. Nearly all cancer cells have some degree of chromosomal instability, but it is unclear whether this drives tumorigenesis or is a by-product of cancer evolution. There is evidence to suggest that a background of chromosomal instability accelerates the rate of tumourigenesis aided by the acquisition of mutations that promote tolerance of instability (for example, inactivation of TP53). Furthermore, the role of recombination activating gene (RAG) and activation-induced cytidine deaminase (AID) in the diversification of immunoglobulin-encoding genes has been proposed as a driver of SVs in leukaemia. The RAG1 and RAG2 endonucleases introduce DSBs followed by the recombination of the immunoglobulin variable, diversity, and joining regions. AID deaminates cytosine residues in the immunoglobulin switch and variable regions, allowing somatic hypermutation. This genetic diversification of antibodies is activated in response to infectious or inflammatory stimuli and has been shown to contribute to the accumulation of somatic mutations necessary for clonal evolution in leukaemia. The mechanisms underlying the generation of structural variation outlined above have been reviewed extensively elsewhere.

The traditional model of cancer development involves the gradual sequential accumulation of mutations, which leads to gradual progression of normal tissue through increasingly disordered clinical and pathological stages until it becomes a malignant tumour. An alternative model was suggested by Stephens et al. to explain complex patterns of genomic rearrangement. They suggest a catastrophe-like model where multiple SVs occur as a result of a single event.

It is likely that such a model of mutation is behind complex patterns of genomic rearrangements found in cancer, such as chromothripsis and chromoplexy. Chromothripsis is characterised by tens to hundreds of genomic rearrangements across an extended chromosomal region. A region that has undergone chromothripsis shows the following hallmarks: oscillating copy number between usually two copy number states; the retention of heterozygosity in the regions with the higher copy number state; and clustering of SV breakpoints to a greater extent than would be expected given a background model of the regional propensity to rearrangement. Chromothripsis appears to be present in 2% to 3% of all cancers and has a higher prevalence in bone cancers (25%) (release 27). It has been reported to be associated with the presence of TP53 mutations in tumours of diverse origins, including medulloblastomas, acute myeloid leukaemia, and pancreatic tumours, suggesting a connection to general genome instability. Moreover, chromothripsis has been associated with poorer prognosis in neuroblastoma, acute myeloid leukaemia, multiple myeloma, and malignant melanoma. In contrast, it has been linked with longer progression-free survival in colorectal cancer. This suggests that chromothripsis has the potential to be an informative biomarker in tumours but with implications that may differ by cancer type. A further complex pattern of structural rearrangements is chromoplexy. This is characterised by chains of translocations and deletions that are interdependent. It differs from chromothripsis in that it involves fewer rearrangements per chain (3–40+) and the rearrangements can be spread across up to 10 chromosomes. It is thought that multiple chains of chromoplexy can be formed in successive cell cycles but that all of the SVs in each chain are likely to be formed by one event. A further sign of chromoplexy is the presence of small deletions between fusion junctions of some translocation rearrangements known as “deletion bridges.” Chromoplexy is largely copy number neutral and has a high prevalence in prostate cancers (88%).

Rates of structural mutation vary across the genome, and there are some regions of the genome that are more susceptible to recurrent structural mutation than others. Glodzik et al. used a large WGS breast cancer cohort to identify such regions that appear to harbour greater recurrent structural variation than expected given the genomic characteristics of the region. These characteristics included replication timing domains, gene-rich regions, background copy number, chromatin states, and repetitive sequences. Many of the rearrangement hotspots found were dominated by long (greater than 100 kb) tandem duplications. These hotspots often duplicated key driver genes and regulatory elements and were enriched for breast cancer susceptibility loci and tissue-specific super enhancers. Thus, based upon the functional regions included, some fraction of these hotspots may represent driver events under selection in tumours, but many hotspots may also be generated by mutational bias at these regions (such that they are more prone to mutation or less accessible to repair) mediated by their chromatin structure and replication timing. Regardless of the relative roles of selection or other processes in their origins, as we discuss below, these patterns of recurrent structural mutation may be clinically informative.

From sequence to data to treatment

The role of genomics in precision therapy often focuses on using targeted sequencing technologies to test for the presence of clinically relevant SNVs. However, certain tumour types are characterised by modest SNV loads and instead show extensive structural rearrangements. Indeed, Ciriello et al. suggest that there is an inverse relationship between the burden of SNVs and CNAs across tumour types (Figure 2a of 82). Figure 1 illustrates the average number of SVs per patient observed across 19 projects from ICGC (release 27). Although it can be challenging to adjust for the inter-study differences highlighted below, it appears that certain (e.g. ovarian, breast, and prostate) tumours possess high frequencies of SVs but that other tumour types such as bone cancers and chronic lymphocytic leukaemia do not. However, there are also noticeable differences in the frequencies of mutations reported between cohorts of the same tumour type, which may reflect differences in cellularity, sequencing depth, and the SV callers used across studies. These differences emphasise the challenges that exist in measuring the degree of structural variation across studies with multiple confounding factors. Nevertheless, the clinical utility of certain SVs is already established. Currently, clinically actionable SVs include gene fusions (for example, BCR-ABL), oncogene amplifications (for example, ERBB2), and tumour
Figure 1. Average number of structural variants per patient by cancer type. The average number of structural variants per patient varies across 19 whole genome sequencing studies of cancer from the International Cancer Genomics Consortium (ICGC) (release 27). Structural variants were classified into five types: deletions, insertions, tandem duplications, inversions, and translocations. Some tumours such as ovarian, breast, and prostate have relatively large numbers of structural variants, whereas bone cancers and chronic lymphocytic leukaemias have relatively few. However, biological differences among tumour types can be confounded by technical differences among studies.

SV, structural variant.

Suppressor deletions (for example, BRCA1). Mertens et al. discuss the impact of gene fusions in cancer and describe their utility as effective drug targets (Table 2 of 85). Further advances in technology will be necessary for the reliable identification of more complex SV patterns in a clinical setting where low-cost and rapid-turnaround assays are desirable.

In the past, the analysis of somatic mutations in cancer has often focused on recurrent driver mutations located within the small fraction of the genome encoding proteins and thought to confer selective advantages to tumour cells. However, it has been convincingly demonstrated that the total burden of mutations genome-wide in a tumour can represent the cumulative impact of mutagens, repair processes, and other influences along the path of tumourigenesis. This overall burden of mutation can be disentangled to identify distinct and independent patterns of mutation which represent distinct mutational processes. The associations seen between these patterns and particular DNA repair deficiencies were recently validated in vitro by using isogenic cell models. The computational analyses required to disentangle these patterns have been developed by using large pan-cancer cohorts of WGS/whole exome sequencing as well as for a large cohort of breast cancer WGS and have successfully identified over 30 patterns or signatures of SNVs classified by their substitution type and trinucleotide context. Analogous approaches have been proposed for the analyses of indels and larger structural rearrangements; in the latter, SVs are classified on the basis of whether or not their breakpoints fall in a cluster of breakpoints in the genome. The rationale behind this separation is that clusters of breakpoints may be formed by mutational processes that are different from those that form dispersed breakpoints. SVs are also classified by their type and size, and compound mutational signatures, including SNVs and SV breakpoints, can be constructed. This approach was successfully tested by using a large cohort of breast cancer WGS, resulting in the identification of characteristic signatures of rearrangement, reflecting HR deficiency, associated with the inactivating mutations of the BRCA1/2 genes. This signature of HR deficiency could then be sought in other tumour samples lacking mutated BRCA1/2 genes by using a new algorithm: HRDetect. This significantly expands the number of patients who have tumours showing HR deficiency (for example, due to epigenomic inactivation of BRCA1/2 or mutations in other HR genes) and who are expected to respond to therapies exploiting defective repair, such as poly (ADP-ribose) polymerase (PARP) inhibitors. The authors also demonstrated the success of HRDetect in predicting HR deficiency in ovarian and pancreatic tumours, suggesting broader potential for therapeutic stratification of patients. Another study has since validated these results, showing that HRDetect provides clinically relevant information independently of BRCA1/2 mutation status and that high HRDetect scores identified patients with good responses to platinum-based chemotherapy. However, the input to HRDetect relies on SV calls from a single SV caller. The incorporation of
variant calls from additional callers to form consensus calls may be beneficial in ensuring the tool’s robustness. Beyond patient stratification, it has also been shown that SV mutational signatures can be used to predict overall survival in a cohort of HGSOC. Finally, it has been shown that the burden of SVs and SNVs in a tumour is associated with the degree of infiltration by host immune cells, linking the mutational state of a tumour to its tendency to elicit an immune response. At the cutting edge of cancer immunotherapy, efforts are under way to exploit such immunogenic tumour mutations and design individual anti-cancer vaccines, placing tumour genomics at the centre of oncology.

The last few decades have transformed our view of cancer to reveal each tumour as a largely unique constellation of SNV and SV mutations, and a tiny fraction of these mutations drive tumourigenesis. The complete reconstruction of the evolutionary trajectory travelled by a tumour could provide new biomarkers for rapid diagnosis and effective treatment and ultimately allow us to predict cancer progression. This comprehensive understanding is currently out of reach and, given the challenges involved, may remain so in the near future. However, in spite of our ignorance about the functional impacts of SVs and our modest success in detecting somatic mutations accurately, there are good reasons for optimism. Although we still lack the sample sizes required to detect all driver genes for most cancer types, there have already been unexpected benefits in amassing WGS tumour cohorts, with the genome-wide patterns of damage seen in a tumour providing novel biomarkers and guiding treatment.

The next decade will see an explosion in the available data and improvements in analysis, making genomics a routine component of medical practice in oncology.

**Abbreviations**

AID, activation-induced cytidine deaminase; CML, chronic myeloid leukemia; CNA, copy number alteration; DSB, double-strand break; HGSOC, high-grade serous ovarian cancer; HR, homologous recombination; ICGC, International Cancer Genome Consortium; NHEJ, non-homologous end joining; RAG, recombination activating gene; SNV, single-nucleotide variant; SV, structural variant; TCGA, The Cancer Genome Atlas; WGS, whole genome sequencing.

**Data and software availability**

The data used to generate Figure 1 are available from the ICGC data portal at the following link: https://dcc.icgc.org/releases/current/Projects.

**Grant information**

The authors are supported by core funding of the UK Medical Research Council (MRC) to the MRC Human Genetics Unit and a UKRI Innovation Fellowship to Allith Ewing.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

---

**References**

1. Inaki K, Liu ET: Structural mutations in cancer: mechanistic and functional insights. Trends Genet. 2012; 28(11): 550–9. [PubMed Abstract](#) [Publisher Full Text](#)
2. Guan P, Sung WK: Structural variation detection using next-generation sequencing data: A comparative technical review. Methods. 2016; 102: 36–49. [PubMed Abstract](#) [Publisher Full Text](#)
3. Yang L, Luquette LJ, Gehlenborg N, et al.: Diverse mechanisms of somatic structural variations in human cancer genomes. Cell. 2013; 153(4): 919–29. [PubMed Abstract](#) [Publisher Full Text](#)
4. Bostell DB, Böhm S, Ahmed AA, et al.: Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer. Nat Rev Cancer. 2015; 15(11): 668–79. [PubMed Abstract](#) [Publisher Full Text](#) [Free Full Text](#)
5. Patch AM, Christie EL, Etemadmoghadam D, et al.: Whole-genome characterization of chemoresistant ovarian cancer. Nature. 2015; 521(7553): 489–94. [PubMed Abstract](#) [Publisher Full Text](#) [F1000 Recommendation](#)
6. Scapa A, Chang DK, Nones K, et al.: Whole-genome landscape of pancreatic neuroendocrine tumours. Nature. 2017; 543(7643): 65–71. [PubMed Abstract](#) [Publisher Full Text](#)
7. Tubio JM: Somatic structural variation and cancer. Brief Funct Genomics. 2015; 14(5): 339–51. [PubMed Abstract](#) [Publisher Full Text](#)
8. Liu B, Conroy JM, Morrison CD, et al.: Structural variation discovery in the cancer genome using next generation sequencing: computational solutions and perspectives. Oncotarget. 2015; 6(8): 5477–89. [PubMed Abstract](#) [Publisher Full Text](#) [Free Full Text](#)
9. Liu B, Morrison CD, Johnson CS, et al.: Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. Oncotarget. 2013; 4(11): 1868–81. [PubMed Abstract](#) [Publisher Full Text](#) [Free Full Text](#)
10. Beroukhim R, Mermel CH, Porter D, et al.: The landscape of somatic copy-number alteration across human cancers. Nature. 2010; 463(7283): 899–905. [PubMed Abstract](#) [Publisher Full Text](#) [Free Full Text](#)
11. Shlien A, Malkin D: Copy number variations and cancer. Genome Med. 2009; 1(6): 62. [PubMed Abstract](#) [Publisher Full Text](#) [Free Full Text](#)
12. Albertson DG, Collins C, McCormick F, et al.: Chromosome aberrations in solid tumors. Nat Genet. 2003; 34(4): 369–76. [PubMed Abstract](#) [Publisher Full Text](#)
13. Dison SJ, Hou C, Gissner JT, et al.: Copy number variation at 1q21.1 associated with neuroblastoma. Nature. 2009; 459(7249): 987–91. [PubMed Abstract](#) [Publisher Full Text](#) [Free Full Text](#)
14. McGranahan N, Rosenthal R, Hiley CT, et al.: Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. Cell. 2017; 171(6): 1259–1271.e11. [PubMed Abstract](#) [Publisher Full Text](#) [Free Full Text](#)
15. Natrajan R, Wilkerson PM, Marchiò C, et al.: Characterization of the genomic features and expressed fusion genes in micropapillary carcinomas of the breast. J Pathol. 2014; 232(3): 563–65. [PubMed Abstract](#) [Publisher Full Text](#) [Free Full Text](#)
16. Wu YM, Su F, Kalyana-Sundaram S, et al.: Identification of targetable FGFR gene fusions in diverse cancers. Cancer Discov. 2013; 3(6): 636–47. [PubMed Abstract](#) [Publisher Full Text](#) [Free Full Text](#)
17. Karlsson J, Liljebjörn H, Holmqvist Mengebror L, et al.: Activation of human telomerase reverse transcriptase through gene fusion in clear cell sarcoma of the kidney. Cancer Lett. 2015; 357(2): 498–501. [PubMed Abstract](#) [Publisher Full Text](#) [Free Full Text](#)
18. Giacomini CP, Sun S, Varma S, et al.: Breakpoint analysis of transcriptional and genomic profiles uncovers novel gene fusions spanning multiple human cancer types. PLoS Genet. 2013; 9(4): e1003464. [PubMed Abstract](#) [Publisher Full Text](#) [Free Full Text](#)
41. Wang J, Mullighan CG, Easton J, et al.: Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. Nat Med. 2017; 23(12): 1646–51. 
Published Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

42. Rausch T, Zichner T, Schlatl A, et al.: DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012; 28(18): 2333–39. 
Published Abstract | Publisher Full Text | Free Full Text

43. Lemire AD, Houlihan KE, Hu Y, et al.: Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat Methods. 2015; 12(7): 623–30. 
Published Abstract | Publisher Full Text | Free Full Text

44. Chong Z, Ruan J, Gao M, et al.: novoBreak: a novel assembly for breakpoint detection in cancer genomes. Nat Methods. 2017; 14(1): 65–72. 
Published Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

45. Guan P, Sung WK: Structural variation detection using next-generation sequencing data: A comparative technical review. Methods. 2016; 102: 36–49. 
Published Abstract | Publisher Full Text

46. Escaramis G, Dicamopo E, Rabionet R: A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics. 2015; 14(5): 305–14. 
Published Abstract | Publisher Full Text

47. Eilott K, Bailey MH, Saksena G, et al.: Scalable Open Sequence Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst. 2018; 6(3): 271–281.e7. 
Published Abstract | Publisher Full Text | Free Full Text

48. Li Y, Roberts N, Weischenfeldt J, et al.: Patterns of structural variation in human cancer. bioRxiv. 2017. 
Published Full Text

49. Holmann AL, Behr J, Singer J, et al.: Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers. BMC Bioinformatics. 2017; 18(1): 8. 
Published Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

50. Mägi A, Sameralo R, Rinnguo A, et al.: NanoPore sequencing data analysis: state of the art, applications and challenges. Brief Bioinform. 2017; box062. 
Published Abstract | Publisher Full Text

51. Currall BB, Chiang C, Tawalkar ME, et al.: Mechanisms for Structural Variation in the Human Genome. Curr Genet Med Rep. 2013; 1(2): 81–90. 
Published Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

52. Stephens PJ, McBride DJ, Lin ML, et al.: Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature. 2009; 462(7276): 1005–10. 
Published Abstract | Publisher Full Text

53. Carvalho CM, Lupsar JR: Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016; 17(4): 224–38. 
Published Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

54. Stephens PJ, Greenman CD, Fu B, et al.: Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell. 2011; 144(1): 27–40. 
Published Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

55. Lee JA, Carvalho CM, Lupsar JR: A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell. 2007; 131(7): 1235–47. 
Published Abstract | Publisher Full Text

56. Hastings PJ, Ira G, Lupsar JR: A microhomology-mediated break-induced replication model for the origin of human copy number variation. PloS Genet. 2009; 5(11): e1000327. 
Published Abstract | Publisher Full Text | Free Full Text

57. Zhang F, Khajavi M, Connolly AM, et al.: The DNA replication FosStes/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat Genet. 2005; 37(7): 849–53. 
Published Abstract | Publisher Full Text | Free Full Text

58. Lee JK, Choi YL, Kwon M, et al.: Mechanisms and Consequences of Cancer Genomic Instability: Lessons from Genome Sequencing Studies. Annu Rev Pathol. 2016; 11: 283–312. 
Published Abstract | Publisher Full Text

59. Swaminathan S, Klemm L, Park E, et al.: Mechanisms of clonal evolution in childhood acute lymphoblastic leukemia. Nat Immunol. 2015, 16(7): 766–74. 
Published Abstract | Publisher Full Text | Free Full Text

60. Papaemmanuil E, Rapado I, Li Y, et al.: RAG-mediated recombination is the predominant driver of oncogenic rearrangement in TETV-RUNX1 acute lymphoblastic leukemia. Nat Genet. 2014; 46(2): 116–25. 
Published Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

61. YI K, JA K: Patterns and mechanisms of structural variations in human cancer. Exp Mol Med. 2018; 50(8): 98. 
Published Abstract | Publisher Full Text | Free Full Text

62. de Pagter MS, Kloosterman WP: The Diverse Effects of Complex Chromosome Rearrangements and Chromosome Instability in Cancer Development. Recent Results Cancer Res. 2015; 200: 165–93. 
Published Abstract | Publisher Full Text

63. Holland AJ, Cleveland DW: Chromosome rearrangement and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. Nat Med.
Precision Therapeutics. Trends Genet. 2016; 32(9): 530–42.

84. International Cancer Genome Consortium, Hudson TJ, Anderson W, et al.: International network of cancer genome projects. Nature. 2010; 464(7291): 993–8.

85. Mertens F, Johannsson B, Fioretos T, et al.: The emerging complexity of gene fusions in cancer. Nat Rev Cancer. 2015; 15(6): 371–81.

86. Stratton MR, Campbell PJ, Futreal PA, et al.: The cancer genome. Nature. 2009; 458(7239): 719–24.

87. Stratton MR: Exploring the genomes of cancer cells: progress and promise. Science. 2011; 331(6024): 1553–8.

88. Nil-Zainaal S, Alexandrov LB, Wedge DC, et al.: Mutational processes molding the genomes of 21 breast cancers. Cell. 2012; 148(6): 579–93.

89. Nil-Zainaal S, Morganella S: Mutation Signatures in Breast Cancer: The Problem at the DNA Level, Clin Cancer Res. 2017; 23(11): 2617–29.

90. Alexandrov LB, Nil-Zainaal S, Wedge DC, et al.: Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 2013; 3(1): 248–65.

91. Alexandrov LB, Nil-Zainaal S, Wedge DC, et al.: Signatures of mutational processes in human cancer. Nature. 2013; 500(7463): 415–21.

92. Zhou X, Ouwus M, Harris R, et al.: Validating the concept of mutational signatures with isogenic cell models. Nat Commun. 2018; 9(1): 1744.

93. Hillman RT, Chisholm GB, Lu KH, et al.: COSMIC: Somatic cancer genetics at high-resolution. Nucl Acids Res. 2017; 45(1): D777–D783.

94. Forbes SA, Beare D, Boutsalakis H, et al.: Genomics in Cancer: The emerging complexity of gene expression. Nat Rev Genet. 2017; 18(10): 547–59.

95. Davies H, Glodzik D, Morganella S, et al.: HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat Med. 2017; 23(4): 517–25.

96. Zhao EY, Shen Y, Pleasance E, et al.: Deciphering signatures of mutational processes in human cancer. Science. 2013; 340(6139): 1029–1041.e21.

97. Helleday T, Eshat S, Nil-Zainaal S: Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet. 2014; 15(9): 585–98.

98. Ouwus M, Harris R, et al.: Deciphering signatures of mutational processes in human cancer. Nature. 2013; 500(7463): 415–21.

99. Hillman RT, Chisholm GB, Lu KH, et al.: Genomic Rearrangement Signatures and Clinical Outcomes in High-Grade Serous Ovarian Cancer. J Natl Cancer Inst. 2018; 110(3): 265–272.

100. Thorsson V, Gibbs DL, Brown SD, et al.: The immune landscape of cancer. Science. 2018; 362(6417): 1231–40.

101. Berger MF, Mardis ER: The emerging clinical relevance of genomics in cancer medicine. Nat Rev Clin Oncol. 2018; 15(6): 353–65.

102. Martincicora I, Raine KM, Gerstung M, et al.: Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017; 171(5): 1029–1041.e21.

103. Jolly C, van Loo P: Timing somatic events in the evolution of cancer. Genome Biol. 2018; 19(1): 95.

104. Lawrence MS, Stojanov P, Mermel CH, et al.: Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014; 500(7484): 495–501.
Open Peer Review

Current Referee Status: ✅ ✅

Editorial Note on the Review Process

F1000 Faculty Reviews are commissioned from members of the prestigious F1000 Faculty and are edited as a service to readers. In order to make these reviews as comprehensive and accessible as possible, the referees provide input before publication and only the final, revised version is published. The referees who approved the final version are listed with their names and affiliations but without their reports on earlier versions (any comments will already have been addressed in the published version).

The referees who approved this article are:

Version 1

1 Javier Herrero Bill Lyons Informatics Centre, UCL Cancer Institute, University College London, London, WC1E 6BT, UK
   Competing Interests: No competing interests were disclosed.
2 Jianmin Wang Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA
   Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com