Published in final edited form as:

Curr Opin Syst Biol. 2017 February ; 1: 54–61. doi:10.1016/j.coisb.2016.12.005.

Modeling cancer rearrangement landscapes

John Maciejowski^{1,2} and Marcin Imielinski^{2,3}

¹Rockefeller University, New York, USA

²New York Genome Center, New York, USA

³Weill Cornell Medicine, New York, USA

Abstract

Cancer genome sequences contain footprints of somatic mutational processes, whose analysis in large tumor sequencing datasets has revealed novel mutational signatures, correlative features of variant topography, and complex events. Many of these analytic results have yet to be reconciled with decades of mechanistic genome integrity research performed in controlled model systems. However, a new generation of genome-integrity experiments combining computational modeling, data analytics, and high-throughput sequencing are emerging to link mechanisms to patterns. Conversely, analytic studies evaluating quantitative footprints of specific genome integrity hypotheses will be critical in fitting naturally occurring mutational patterns to the predictions of a particular mechanistic model. Such quantitative and mechanistic studies will form the foundation of an emerging systems biology of genome integrity.

Keywords

cancer; genomics; rearrangements; genome integrity; next generation sequencing; assembly; tumor evolution

Introduction

Cancer genomic mutational landscapes are shaped by the forces of selection, mutagenesis, and repair. In recent years, massively parallel sequencing has enabled the first unbiased large-scale surveys of somatic genome variation in a variety of cancer types [1]. Though primarily motivated by the search for driver alterations and novel cancer genes, these studies have also revealed (and have been occasionally confounded by) the footprints of known and novel mutational processes [2,3]. These data-driven, statistical analyses provide a fascinating counterpoint to decades of genome integrity research, where rigorous experimentation has dissected the biochemistry and cell biology of DNA damage and repair in model systems such as yeast and mammalian lymphocytes [4].

Though the interaction between cancer genomics and genome integrity fields has been mostly ad hoc, it has already generated several important successes. This includes examples where mechanistic intuition has been statistically validated through cancer genome analyses,

and where cancer genome hypotheses have been modeled in experimental systems. Though the most mature area of research has been the study of single nucleotide variants (SNVs), the biggest challenges and opportunities for interactions between computationalists and experimentalists lie in the analysis of DNA rearrangements. We argue that these interactions form the seeds of an emerging field of the *systems biology of cancer genome integrity* that refines mechanistic models of somatic DNA repair pathways through quantitative analysis of high-throughput sequencing-based signals.

SNV landscapes

The study of single nucleotide variants (SNVs) patterns provides a paradigm for cancer genomics-driven mutational process research [5]. Analytic work in this area has focused on three key directions: mutational signatures, variant topography, and the identification of complex events. As many somatic SNVs are not linked to a known biological mechanism, they provide many fascinating starting points for future experimental investigation.

Somatic SNVs arise under the combined action of exogenous mutagens (e.g. UV light, tobacco carcinogens, industrial chemicals), DNA repair defects (e.g. base and nucleotide excision repair, mismatch repair), endogenous mutational mechanisms (e.g. activation-induced cytidine deaminase), and selection [2]. Cancer genomes bear distinct signatures of these mechanisms, which can be discovered *de novo* through factorization of SNV count matrices cataloguing 192 stranded or 96 strand-collapsed tri-nucleotide sequence contexts (e.g. TpCpA > T, where C is the mutated base) [6]. Analysis of these signatures has been used to identify footprints of aging [7], acquired DNA repair defects in POLE [8] and ERCC2 [9], and mutagenic exposures (e.g. aristolochic acid) [10].

Correlative analyses of variant topography have linked regional fluctuations in SNV density to replication timing [3,11], post-translational chromatin modification [12,13], transcription factor binding [14–16], and tissue-specific chromatin accessibility patterns [17]. At the forefront of SNV mutational processes research are investigations into the regional fluctuations in specific base-context and strand-specific SNV signatures [8,18]. Such approaches have been used to attribute specific mutational processes (e.g. mismatch-repair deficiency) to replication vs transcription-based mechanisms.

Finally, WGS analyses have been used to identify complex mutational events that generate multiple SNVs *in cis*, such as *kataegis*. Kataegis refers to a phenomenon of strand-coordinated clusters of cytosine mutations in the TpCp* context, occurring near somatic rearrangements and affecting specific cancer types (e.g. breast cancer) [19,20]. Kataegis SNV patterns may derive from APOBEC enzyme activity [21–23].

Complex rearrangement landscapes

Cancer DNA rearrangements have been linked to specific mechanisms of DNA breakage and repair, including nonhomologous end joining (NHEJ), microhomology mediated end-joining (MMEJ), DNA replication stress, telomere fusions, and breakage fusion bridge cycles (BFBs) [24]. Though many of these mechanisms have been convincingly established in

model systems (e.g. yeast, mammalian B-cells), their relevance to cancer DNA structural variation has not been comprehensively evaluated in large pan-cancer sequencing datasets.

Rearrangement signature analyses have catalogued spectra of structural variant types (deletion, amplification, translocation) and sequence contexts across multiple cancers, however at a much smaller scale than for SNVs. Though many cancer genomes are highly rearranged, the average genome harbors fewer than a few hundred rearrangements, as opposed to tens of thousands of SNVs [25]. In addition, identification of rearrangements requires WGS data, of which only ~3000 cases have been made available to the community (in contrast to the 15,000 exome cases that have been deposited in data portals). The largest published pan-cancer rearrangement signatures studies have analyzed fewer than 200 samples [25–27]. While some groups have observed short stretches of sequence homology around rearrangement junctions, it is unclear whether the observed patterns provide definitive evidence for alternate rearrangement mechanisms (e.g. MMEJ). Much less is known about the genetic basis of hypermutators for rearrangements than for SNVs, though enrichment of specific rearrangement signatures have been noted in a few tumor contexts: Rag recombinase deletion signature in ETV6-RUNX1 fused acute lymphoblastic lymphomas [28], translocation in exceptional platinum-chemotherapy responder pancreatic cancer [29]. Current international analysis efforts are underway to analyze much larger datasets [30].

The topography of cancer rearrangement patterns has also not been as exhaustively explored as for SNVs. Pan-cancer microarray analyses have correlated copy number endpoint locations with late replicating regions and nuclear proximity as determined by 3D chromatin confirmation (Hi-C) methods [31–33]. Similar analyses have not been, to our knowledge, applied to pan-cancer whole genome sequencing data. However, a recent study analyzing rearrangement patterns in over 500 breast cancer whole genomes has suggested that somatic rearrangement junctions are enriched in early, rather than late, replicating regions [34]. Mechanistic studies incorporating engineered translocations in mouse model systems have suggested that rearrangement junctions occur independently of nuclear proximity [35]. Additional pan-cancer analyses of whole genome sequences will be required to resolve these apparent contradictions, and suggest novel topographic correlations.

Despite the paucity of large-scale statistical surveys, some of the most dramatic findings in cancer genomics have emerged from WGS rearrangement analyses. This includes the discovery of chromothripsis (chromosome shattering [36]) and chromoplexy (chains of balanced rearrangements [37]). The concept of “catastrophic” large-scale rearrangement events has provided an attractive explanation for these phenomena [38], though alternative explanations invoking progressive iteration of simple events have been argued using computational models [39]. Such modeling approaches provide a lens with which to understand the genomic footprints of BFBs, a complex rearrangement phenomenon described long before the sequencing era [40,41]. Modeling becomes especially critical when BFBs or chromothripsis are observed in conjunction with other rearrangement events [42]. It remains to be seen whether the existing “lexicon” of rearrangement events are sufficient to explain the dense webs of rearrangement junctions observed in many cancer sequence data, or whether additional basic event types remain to be discovered [43].

Fundamental analytic challenges

Several conceptual and data-driven challenges complicate the analysis of rearrangement patterns in cancer genomes: As mentioned above, rearrangement data are sparse, comprising much fewer events per sample than for SNVs. Additional complexity is related to the definition of a rearrangement “event” and rearrangement context. Finally, the myopic nature of short-read sequencing (100–200 bp read pairs generated from 300 to 700 bp DNA fragments), makes it difficult to assemble or phase rearranged loci and detect variants in repetitive genomic regions [44].

The atomic unit of cancer rearrangement data is a *junction* (Figure 1): a pair of genomic locations and strands that were previously distant but were made adjacent through a rearrangement *event*. There are four orientations that a junction can take with respect to a particular coordinate system, which is most commonly defined with respect to the *p* and *q* arm orientation of chromosomes in a reference assembly, however could also be defined with respect to other frames (e.g. replication origin, transcription origin). A rearrangement event may involve a single junction or multiple junctions, e.g. a deletion or duplication involves a single junction, a balanced translocation or inversion involve two junctions, chromothripsis or chromoplexy involve many more [38].

When multiple junctions map to a given region, determining the events that gave rise to them may not be straightforward. As Figure 2 illustrates, a junction that appears to be a deletion locally (inner junction associated with copy number drop) is more plausibly explained as a duplication when the broader context is considered (outer junction associated with copy number gain). For this reason, the correct inference of rearrangement events from WGS junction patterns requires building models of loci evolving across intermediate and final states. When many junctions map nearby on the genome, this inference may be computationally intensive and fundamentally ambiguous, with many solutions fitting the data equally well. Resolution of this ambiguity may be facilitated with additional data, either snapshots of additional tumor samples in intermediate temporal states of evolution or long-range sequencing data enabling end-to-end locus assembly in space. Technologies providing long-range *cis* information include single molecule sequencing (e.g. Pacific Biosciences), optical mapping (e.g. BioNano genomics), linked reads (e.g. 10X genomics), and proximity ligation (e.g. Dovetail genomics) [45–47]. The development and application of novel computational modeling approaches [48,49], based partly on phylogenetic techniques employed in comparative genomics, may provide additional inferential power for reconstructing rearranged loci.

In the absence of perfect reconstructions of locus evolution or *cis*-structure, it can still be possible to study sequence signatures or topography with respect to individual junctions (Figure 3). Unlike with SNVs, the sequence and topographic context that must be mined must be gleaned from two or more genomic locations. Instead of simply counting base triplets, sequence context analysis must examine stretches of homology at or near both junction ends, and take into account additional somatic SNVs occurring near the lesion site (e.g. TpC mutations arising in the context of kataegis). This sequence context analysis must be oriented with respect to the junction strands, and also take into account (novel or

templated) sequence that may be inserted between the junction ends. Similar consideration must be applied to the analysis of topographic genomic features, such as regional chromatin states or other features (e.g. R-loops) that are gleaned from “*-seq” experiments. A final feature of rearrangement context is the proximity between junction ends, which can be measured using chromatin conformation maps. These analyses can be used to examine relationships between the context (the state of the genome prior to an event) and consequence (the state of the genome following the event) of a somatic junction.

From pattern to mechanism, and back

The identification of recurrent patterns in genomic data inspires the search for mechanisms, through the construction of experimental systems and conceptual models that are sufficient to generate a set of observations. The prototypical example of this link between pattern and mechanism has been the study of chromothripsis, where elegant engineering of cells in culture have established sufficient conditions for the phenomenon of “chromosome shattering”. Zhang and colleagues employed live cell microscopy and single-cell sequencing to demonstrate a high frequency of chromothripsis induction in the immediate progeny of cells with induced micronuclei [50]. Maciejowski and colleagues induced chromosomal bridges via telomere dysfunction and demonstrated chromothripsis and kataegis through WGS of derived clones [51]. Mardin and colleagues employed chemical double stranded break induction and selection for anchorage independence growth followed by WGS of clones to recapitulate chromothripsis in culture [52]. While these studies have effectively solved a prominent cancer genomics puzzle, they also have also spawned new ones: What is the contribution of these competing mechanisms to chromothripsis events observed in real tumors? Can their signatures be recognized and analytically disentangled using cancer sequencing data?

The approach taken with the modeling of chromothripsis provides a template with which to mechanistically dissect other complex rearrangement phenomena, whether chromoplexy or BFB cycles. Such experiments may be proposed to follow a common template (Figure 4) following classic reverse-genetics approaches but employing modern genomic-profiling technology and analytic tools derived from machine learning. Such an approach has been employed to study BFBs in *Caenorhabditis elegans*, where whole genome sequencing of strains defective in telomere maintenance revealed clusters of rearrangements near chromosome ends that were consistent with the proposed BFB signature [53]. It will be important to explore similar experimental avenues in mammalian systems as holocentric *C. elegans* chromosomes can be stably maintained after fusion events and it is not clear how accurately the rearrangements observed in this system can explain rearrangement signatures observed in cancer genomics. Epithelial cancers that develop after telomere dysfunction in mice show chromosome rearrangements with amplified regions that are adjacent to translocations [54]. This landmark experiment was performed before the advent of massively parallel whole genome sequencing and it would be informative to revisit it and other classic experiments with an eye towards genomic phenotypes in order to conclusively link a major observed feature of cancer genomes with a biological mechanism. The engineered variant patterns generated in these carefully controlled experimental systems could then be analyzed in cancer genome data to close the loop and validate a mechanistic hypothesis.

Towards a systems biology of genome integrity

A systems biology approach to cancer genome integrity offers insights into the mechanisms that maintain genome integrity and into the biology underlying structural DNA variations. We propose an experimental framework that takes the shape of classic reverse genetic approaches to probe the genomic consequences of chemical or genetic perturbations (Figure 4). This biology-first approach is necessary to circumvent the limitations associated with correlative analyses performed in most bioinformatics-based studies, such as deficiencies in genetic pathways that are not under immediate consideration. Implementation of this framework depends on (1) the selection of a genetically amenable or chemically accessible model to be studied. Model systems include cultured cells or widely used model organisms such as mice or worms.

One important consideration is that the system selected for study has a stable genome. The model system will serve as a control in downstream sequencing-based readouts. For the purpose of structural variant-based studies (2) a genetically or chemically engineered biological context can be engineered to test a specific hypothesis. Interesting chemical perturbations include treatment with chemotherapies of uncertain genomic effect such as ionizing radiation or conventional or next generation anti-mitotic agents such as paclitaxel and Mps1 inhibitors. Genetic manipulations can be introduced to determine the proposed dependency on APOBECs for kataegis-associated mutations [21] or BRCA1 deficiency as the mechanistic cause of the tandem duplicator phenotype [55]. In past work, it has been crucial to perform (3) initial profiling in order to focus sequencing efforts on targets with notable structural variants. Prior profiling has included time-lapse microscopy to target chromosome mis-segregation events [50], karyotype screening to identify cell clones with interesting rearrangements [51], and cellular transformation [52]. Finally, (4) genomic and epigenomic readouts of junctions, chromatin states, and breaks will provide contextual information necessary to build structural variant signatures. Breakpoint junctions detected by whole genome sequencing can be analyzed in the context of nuclear proximity in parallel Hi-C assays. Other sequencing-oriented applications such as END-seq can be used to analyze the genomic factors that influence choice of DNA repair pathway [56].

Conclusion

In this review, we have highlighted key results from cancer genomics mutational process research, while advocating a more formal interaction between cancer genomics and genome integrity fields. We have pinpointed several prominent genome integrity hypotheses that have not yet been addressed in cancer genomics data and cancer genome phenomena that have yet to be modeled in experimental systems. Looking forward, we mention specific experimental and algorithmic approaches that could bolster and enrich this interaction, as experimentalists and theoreticians, computationalists and bench biologists join forces to make sense of a fundamental cancer hallmark.

Acknowledgments

John Maciejowski is supported by a grant from the NCI (K99CA212290). Marcin Imielinski is supported by a Burroughs Wellcome Fund Career Award for Medical Sciences. The authors thank Peter Campbell and Yilong Li for conceiving the example behind Figure 2 and allowing us to incorporate the example into this manuscript.

References

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Garraway LA, Lander ES. Lessons from the cancer genome. *Cell*. 2013; 153:17–37. [PubMed: 23540688]
2. Roberts SA, Gordenin DA. Hypermutation in human cancer genomes: footprints and mechanisms. *Nat Rev Cancer*. 2014; 14:786–800. [PubMed: 25568919]
3. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013; 499:214–218. [PubMed: 23770567]
4. Jeggo PA, Pearl LH, Carr AM. DNA repair, genome stability and cancer: a historical perspective. *Nat Rev Cancer*. 2016; 16:35–42. [PubMed: 26667849]
5. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet*. 2014; 15:585–598. [PubMed: 24981601]
6. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg Å, Børresen-Dale A-L, et al. Signatures of mutational processes in human cancer. *Nature*. 2013; 500:415–421. [PubMed: 23945592]
7. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, Stratton MR. Clock-like mutational processes in human somatic cells. *Nat Genet*. 2015; 47:1402–1407. [PubMed: 26551669]
- 8•. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature*. 2015; 521:81–84. The authors analytically demonstrate that somatic SNV contexts associated with mismatch repair deficiency are no longer correlated with replication timing and gene expression. This analysis is among the first to apply a combined analysis of signatures and topography to identify features of somatic mutational processes. [PubMed: 25707793]
9. Kim J, Mouw KW, Polak P, Braunstein LZ, Kamburov A, Tiao G, Kwiatkowski DJ, Rosenberg JE, Van Allen EM, D'Andrea AD, et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet*. 2016; 48:600–606. [PubMed: 27111033]
10. Petljak M, Alexandrov LB. Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis*. 2016; 37:531–540. [PubMed: 27207657]
11. Woo YH, Li W-H. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat Commun*. 2012; 3:1004. [PubMed: 22893128]
12. Hodgkinson A, Chen Y, Eyre-Walker A. The large-scale distribution of somatic mutations in cancer genomes. *Hum Mutat*. 2012; 33:136–143. [PubMed: 21953857]
13. Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*. 2012; 488:504–507. [PubMed: 22820252]
14. Polak P, Lawrence MS, Haugen E, Stoletzki N, Stojanov P, Thurman RE, Garraway LA, Mirkin S, Getz G, Stamatoyannopoulos JA, et al. Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat Biotechnol*. 2014; 32:71–75. [PubMed: 24336318]
15. Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, Lopez-Bigas N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature*. 2016; 532:264–267. [PubMed: 27075101]

16. Perera D, Poulos RC, Shah A, Beck D, Pimanda JE, Wong JWH. Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature*. 2016; 532:259–263. [PubMed: 27075100]
17. Polak P, Karlic R, Koren A, Thurman R, Sandstrom R, Lawrence MS, Reynolds A, Rynes E, Vlahovick K, Stamatoyannopoulos JA, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*. 2015; 518:360–364. [PubMed: 25693567]
- 18•. Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, Rheinbay E, Kim J, Maruvka YE, Braunstein LZ, et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell*. 2016; 164:538–549. The authors analyze SNV strand asymmetry across many human cancers, and use these features to associate specific mutational context with replication-based vs. transcription-based mechanisms. This analysis represents the forefront of SNV cancer genome analyses, combining both sequence based and topographic features to establish footprints of mutational processes. [PubMed: 26806129]
19. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*. 2012; 149:979–993. [PubMed: 22608084]
20. Roberts SA, Sterling J, Thompson C, Harris S, Mav D, Shah R, Klimczak LJ, Kryukov GV, Malc E, Mieczkowski PA, et al. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol Cell*. 2012; 46:424–435. [PubMed: 22607975]
21. Burns MB, Lackey L, Carpenter MA, Rathore A, Land AM, Leonard B, Refsland EW, Kotandeniya D, Tretyakova N, Nikas JB, et al. APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature*. 2013; 494:366–370. [PubMed: 23389445]
22. Burns MB, Temiz NA, Harris RS. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat Genet*. 2013; 45:977–983. [PubMed: 23852168]
23. Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP, Kim J, Kwiatkowski DJ, Fargo DC, Mieczkowski PA, et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet*. 2015; 47:1067–1072. [PubMed: 26258849]
24. Kass EM, Moynahan ME, Jasin M. When genome maintenance goes badly awry. *Mol Cell*. 2016; 62:777–787. [PubMed: 27259208]
25. Yang L, Luquette LJ, Gehlenborg N, Xi R, Haseley PS, Hsieh C-H, Zhang C, Ren X, Protopopov A, Chin L, et al. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*. 2013; 153:919–929. [PubMed: 23663786]
26. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, Getz G. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res*. 2013; 23:228–235. [PubMed: 23124520]
27. Malhotra A, Lindberg M, Faust GG, Leibowitz ML, Clark RA, Layer RM, Quinlan AR, Hall IM. Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Res*. 2013; 23:762–776. [PubMed: 23410887]
- 28•. Papaemmanuil E, Rapado I, Li Y, Potter NE, Wedge DC, Tubio J, Alexandrov LB, Van Loo P, Cooke SL, Marshall J, et al. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nat Genet*. 2014; 46:116–125. The authors demonstrate a sequence signature of Rag recombination in ETV6-RUNX1 rearranged acute lymphoblastic leukemias. This analysis provides the clearest example of a tumor-type specific rearrangement signature, exploring several dimensions of rearrangement context (local sequence features, regional chromatin features) to characterize the pattern. [PubMed: 24413735]
29. Bailey P, Miller D, Quek K, Quinn MCJ, Robertson AJ, Bruxner TJC, Christ AN, Harliwong I, Manning S, Nourbakhsh E, et al. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*. 2015; 518:495–501. [PubMed: 25719666]
30. Habermann N, Mardin BR, Yakneen S, Korbel JO. Using large-scale genome variation cohorts to decipher the molecular mechanism of cancer. *C R Biol*. 2016; 339:308–313. [PubMed: 27342254]
31. De S, Michor F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat Biotechnol*. 2011; 29:1103–1108. [PubMed: 22101487]

32. De S, Michor F. DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nat Struct Mol Biol.* 2011; 18:950–955. [PubMed: 21725294]
33. Fudenberg G, Getz G, Meyerson M, Mirny LA. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat Biotechnol.* 2011; 29:1109–1113. [PubMed: 22101486]
34. Morganella S, Alexandrov LB, Glodzik D, Zou X, Davies H, Staaf J, Sieuwerts AM, Brinkman AB, Martin S, Ramakrishna M, et al. The topography of mutational processes in breast cancer genomes. *Nat Commun.* 2016; 7:11383. This study comprises the largest published whole genome sequencing analysis of rearrangement signatures and the only known large-scale (>100 sample) survey of rearrangement topography, though limited to breast cancer samples. Among more than 500 cases, they identify 6 rearrangement signatures and suggest significant enrichment of events in early replicating regions. [PubMed: 27136393]
35. Bunting SF, Nussenzweig A. End-joining, translocations and cancer. *Nat Rev Cancer.* 2013; 13:443–454. [PubMed: 23760025]
36. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell.* 2011; 144:27–40. [PubMed: 21215367]
37. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, Park K, Kitabayashi N, MacDonald TY, Ghandi M, et al. Punctuated evolution of prostate cancer genomes. *Cell.* 2013; 153:666–677. [PubMed: 23622249]
38. Korbel JO, Campbell PJ. Criteria for inference of chromothripsis in cancer genomes. *Cell.* 2013; 152:1226–1236. [PubMed: 23498933]
39. Kinsella M, Patel A, Bafna V. The elusive evidence for chromothripsis. *Nucleic Acids Res.* 2014; 42:8231–8242. [PubMed: 24939897]
40. Zakov S, Kinsella M, Bafna V. An algorithmic approach for breakage-fusion-bridge detection in tumor genomes. 2013
41. Greenman CD, Cooke SL, Marshall J, Stratton MR, Campbell PJ. Modeling the evolution space of breakage fusion bridge cycles with a stochastic folding process. *J Math Biol.* 2015; 72:47–86. [PubMed: 25833184]
42. Li Y, Schwab C, Ryan SL, Papaemmanuil E, Robinson HM, Jacobs P, Moorman AV, Dyer S, Borrow J, Griffiths M, et al. Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia. *Nature.* 2014; 508:98–102. The authors demonstrate childhood ALLs arising in the context of an inherited Robertsonian translocation recurrently arise from a somatic evolution pattern consisting of breakage-fusion-bridge cycles that are followed by a chromothripsis event. This is the first study to associate somatic junction patterns with the sequential evolution of two complex rearrangement events. [PubMed: 24670643]
43. Greenman CD, Pleasance ED, Newman S, Yang F, Fu B, Nik-Zainal S, Jones D, Lau KW, Carter N, Edwards PAW, et al. Estimation of rearrangement phylogeny for cancer genomes. *Genome Res.* 2012; 22:346–361. [PubMed: 21994251]
44. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet.* 2012; 13:36–46.
45. Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastie AR, Marks P, Lee J, Chu C, Lin C, Džakula Ž, et al. A hybrid approach for de novo human genome sequence assembly and phasing. *Nat Meth.* 2016; 13:587–590.
46. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* 2016; 26:342–350. [PubMed: 26848124]
47. Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O, Rausch T, Stütz AM, Stedman W, Anantharaman T, Hastie A, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Meth.* 2015; 12:780–786.
48. Zerbino DR, Ballinger T, Paten B, Hickey G, Haussler D. Representing and decomposing genomic structural variants as balanced integer flows on sequence graphs. *BMC Bioinformatics.* 2016; 17:400. [PubMed: 27687569]

49. Schwarz RF, Trinh A, Sipos B, Brenton JD, Goldman N, Markowitz F. Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comput Biol*. 2014; 10:e1003535. [PubMed: 24743184]
- 50••. Zhang C-Z, Spektor A, Cornils H, Francis JM, Jackson EK, Liu S, Meyerson M, Pellman D. Chromothripsis from DNA damage in micronuclei. *Nature*. 2015; 522:179–184. This study marks the first published example of chromothripsis induction in cell culture. The authors induce micronuclei and use live cell imaging and demonstrate chromothripsis in whole genome sequences of daughter nuclei arising after a single cell division. [PubMed: 26017310]
- 51••. Maciejowski J, Li Y, Bosco N, Campbell PJ, de Lange T. Chromothripsis and kataegis induced by telomere crisis. *Cell*. 2015; 163:1641–1654. The authors generate chromothripsis and kataegis in cell lines by inducing telomere dysfunction. They use whole genome sequencing of post-telomere crisis clones to demonstrate the genomic phenotype. [PubMed: 26687355]
- 52••. Mardin BR, Drainas AP, Waszak SM, Weischenfeldt J, Isokane M, Stütz AM, Raeder B, Efthymiopoulos T, Buccitelli C, Segura-Wang M, et al. A cell-based model system links chromothripsis with hyperploidy. *Mol Syst Biol*. 2015; 11:828. The authors propose “CAST”, an experimental system that they use to generate complex DNA rearrangements using cell perturbations followed by selection. They use the system to generate sufficient conditions for generating chromothripsis in culture. [PubMed: 26415501]
53. Meier B, Cooke SL, Weiss J, Bailly AP, Alexandrov LB, Marshall J, Raine K, Maddison M, Anderson E, Stratton MR, et al. *C. elegans* whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res*. 2014; 24:1624–1636. [PubMed: 25030888]
54. Artandi, SE., Chang, S., Lee, SL., Alson, S., Gottlieb, GJ. Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. *Nature*. 2000. <http://dx.doi.org/10.1093/oxfordjhb/9780199368815.013.44>
55. Menghi F, Inaki K, Woo X, Kumar PA, Grzeda KR, Malhotra A, Yadav V, Kim H, Marquez EJ, Ucar D, et al. The tandem duplicator phenotype as a distinct genomic configuration in cancer. *Proc Natl Acad Sci USA*. 2016; 113:E2373–E2382. [PubMed: 27071093]
56. Canela A, Sridharan S, Sciascia N, Tubbs A, Meltzer P, Sleckman BP, Nussenzweig A. DNA breaks and end resection measured genome-wide by end sequencing. *Mol Cell*. 2016; 63:898–911. [PubMed: 27477910]

Figure 1.

DNA junctions: The four possible types of rearrangement junctions are illustrated, each consisting of a pair of breakpoint locations and orientations. The orientations are defined using a particular coordinate system, which is normally anchored with respect to the p - q axis of a chromosome, but can be anchored to other landmarks (e.g. replication origins, transcription start sites). In the standard frame, the left direction is the direction of decreasing coordinates (e.g. towards the p arm) and the right direction is in the direction of increasing coordinates (e.g. towards the q arm). Junctions ends that connect to the left side of a breakpoint are denoted as “-” and those that connect to the right side are denoted with a “+”. Red lines indicate rearrangements connecting two previously separate genomic loci.

Figure 2.

Rearrangement evolution: Junction patterns in cancer are usually the result of the progression of a locus across multiple states of evolution. The derivation of rearrangement *events* from junction patterns requires modeling their evolution in time and/or their long-range structure in space. The example here depicts the challenge of inferring events from junctions, by depicting two tandem duplication events occurring temporally in sequence (Peter Campbell and Yilong Li, personal communication). The left column shows the evolution of the locus assembly (i.e. string of segments A–E) in somatic coordinates, progressing from its initial state to early and advanced lesions in which the central segments are duplicated. The right column shows the same assembly in reference genome coordinates as a thread of segments joined by rearrangement junctions (purple and red). Simulated read data from rearranged genomes generate coverage patterns on the reference that reflect the copy number of segments A–E. For example, the advanced lesion copy number pattern is consistent with 1, 3, 2, 3, and 1 copies of segments A–E, respectively. Since the purple junction is associated with an intervening copy loss, it would be categorized as a deletion event by most copy number callers. However, given the end-to-end assembly of segments, it is clearly the result of a (late) tandem duplication event.

Figure 3.

Junction context: Several notions of junction context are illustrated here, useful for defining and exploring sequence based signatures and regional genomic topographies that might be associated with rearrangement events. Since junctions connect two loci, the immediate and regional sequence context of each breakpoint must be considered (e.g. immediate microhomology in blue, distant homology in red oriented using junction-centric coordinates (where coordinate 0 is the black-gray interface and black denotes positive coordinates (so the coordinates increase away from the junction in two directions)). In addition to the reference sequence context, one might consider inserted sequence at the junction sites and the locations of additional somatic mutations (e.g. kataegic SNVs at the arrowheads) occurring in *cis* to the rearrangement site as context features. In addition to sequence context, numeric genomic annotations (e.g. conservation signal, GC content, DNA replication timing) or the results of [*]-seq experiments (e.g. ChIP-seq, R-loop annotation, nuclear proximity) may be considered as topographical context features. As has been done for SNVs, these definitions of junction context could be used to define footprints of *mutational processes* driving rearrangements, either in naturally occurring cancers or model systems. These context features could be measured in reference states (i.e. prior to rearrangement) or rearranged genomes (e.g. tumor samples) to assess both causes and consequences of rearrangements. Such an approach has been informally used in previous studies [28] to identify novel rearrangement signatures.

Figure 4.

A map for computational mechanistic genome integrity experiments: Experiments begin with (1) an initial cell, tissue, or organism type which undergoes (2) perturbations to alter biological context, ablate specific pathways, and/or induce mutagenesis through one of several mechanisms. Then, to enable (3) initial profiling and selection, nucleic acids are extracted from bulk tissue or via multiple-displacement amplification from single cells (+/- time lapse microscopy). Alternatively, clones are expanded and selected through transformation or karyotype screening. Finally, (4) comprehensive genomic profiling is performed, using whole genome sequencing (via standard short reads or long-range sequencing technologies) or other genomic and epigenomic profiling approaches. The same profiling can also be performed on the unperturbed cell type. The resulting variant data and their contextual features are analyzed using machine learning algorithms to identify recurrent patterns and link them to their biological context (e.g. pathway ablation). Among the several paths in this map are several recent studies dissecting the mechanistic basis for chromothripsis [50–52].