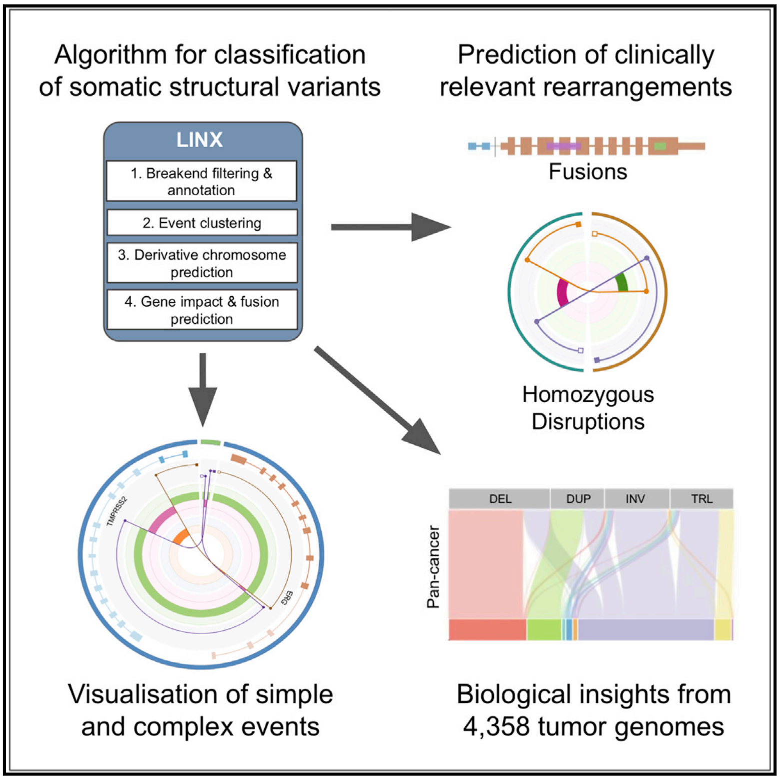
Cell Genomics

Technology

Unscrambling cancer genomes via integrated analysis of structural variation and copy number

Graphical abstract


Algorithm for classification of somatic structural variants
Prediction of clinically relevant rearrangements
LINX
1. Breakend filtering & annotation
2. Event clustering
3. Derivative chromosome prediction
4. Gene impact & fusion prediction
Fusions
Homozygous Disruptions
TMPRSS2
ERG
DEL
DUP
INV
TRL
Pan-cancer
Visualisation of simple and complex events
Biological insights from 4,358 tumor genomes

Highlights

- LINX is an algorithm to classify somatic structural variation in tumors
- Chaining, clustering, and visualizations provide insights into complex rearrangements
- LINX predicts diverse pathogenic rearrangements, including chained fusions
- Homozygous disruptions are a distinct and common driver class in tumors

Authors

Charles Shale, Daniel L. Cameron, Jonathan Baber, ..., Anthony T. Papenfuss, Edwin Cuppen, Peter Priestley

Correspondence

p.priestley@hartwigmedicalfoundation.nl

In brief

Shale et al. present a comprehensive analysis of somatic structural variation in a large cohort of cancer genomes. They have developed an algorithm, LINX, that reveals insights into complex genomic events and demonstrates the utility of whole-genome sequencing in detection of diverse clinically relevant gene fusions and disruptions.

Shale et al., 2022, Cell Genomics *2*, 100112
April 13, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.xgen.2022.100112

Cell Genomics

CellPress
OPEN ACCESS

Technology

Unscrambling cancer genomes via integrated analysis of structural variation and copy number

Charles Shale,[1,2] Daniel L. Cameron,[1,3,4] Jonathan Baber,[1,2] Marie Wong,[5,6] Mark J. Cowley,[5,6]
Anthony T. Papenfuss,[3,4,7,8] Edwin Cuppen,[2,9] and Peter Priestley[1,2,10,*]
[1]Hartwig Medical Foundation Australia, Sydney, NSW, Australia
[2]Hartwig Medical Foundation, Science Park 408, Amsterdam, the Netherlands
[3]Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia
[4]Department of Medical Biology, University of Melbourne, Melbourne, VIC, Australia
[5]Children's Cancer Institute, Lowy Cancer Centre, UNSW Sydney, Kensington, NSW, Australia
[6]School of Women's and Children's Health, UNSW Sydney, Kensington, NSW, Australia
[7]Peter MacCallum Cancer Centre, Melbourne, VIC, Australia
[8]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, VIC, Australia
[9]Center for Molecular Medicine and Oncode Institute, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, the Netherlands
[10]Lead contact
*Correspondence: p.priestley@hartwigmedicalfoundation.nl
https://doi.org/10.1016/j.xgen.2022.100112

SUMMARY

Complex somatic genomic rearrangements and copy number alterations are hallmarks of nearly all cancers. We have developed an algorithm, LINX, to aid interpretation of structural variant and copy number data derived from short-read, whole-genome sequencing. LINX classifies raw structural variant calls into distinct events and predicts their effect on the local structure of the derivative chromosome and the functional impact on affected genes. Visualizations facilitate further investigation of complex rearrangements. LINX allows insights into a diverse range of structural variation events and can reliably detect pathogenic rearrangements, including gene fusions, immunoglobulin enhancer rearrangements, intragenic deletions, and duplications. Uniquely, LINX also predicts chained fusions that we demonstrate account for 13% of clinically relevant oncogenic fusions. LINX also reports a class of inactivation events that we term homozygous disruptions that may be a driver mutation in up to 9% of tumors and may frequently affect *PTEN*, *TP53*, and *RB1*.

INTRODUCTION

Somatic structural variation (SV) and associated copy number alterations (CNAs) are key mechanisms in tumorigenesis.[1] However, both the mechanisms driving and the consequences of genomic rearrangements in cancer are less well understood than for point mutation events. This is due both to the relative paucity of whole-genome sequencing (WGS) data that are required for comprehensive SV analysis and also to the fact that genomic rearrangements have significant diversity. Many rearrangements involve a high degree of complexity, with individual events resulting in multiple or even hundreds of breaks.[2,3] Interpretation of these highly rearranged genomes is challenging but simultaneously highly relevant for the identification of driver events that may function as biomarkers or druggable targets.

LINX is an SV interpretation tool, which integrates CNA and SV calling derived from WGS data and comprehensively clusters, chains, and classifies genomic rearrangements. The motivation for this is twofold: first, from a biological perspective, to allow better insight into distinct mechanisms of rearrangements in tumorigenesis and second, from a clinical perspective, to allow prediction of the functional impact of structural rearrangements, including gene fusions and disruptions. A number of previous tools have been developed to analyze the roles of certain rearrangement event types in tumorigenesis, such as chromothripsis,[2] chromoplexy,[4] long interspersed nuclear element (LINE) insertions,[5] and amplification mechanisms.[6] Clustering methodologies have also been used previously to propose signatures of structural rearrangement.[1,7] LINX goes further than just integrating the functionality of each of these previous tools, both by classifying all classes of rearrangements in each tumor genome and by predicting the local chained structure of the derivative chromosome as well as the functional impact of the rearrangement in a single application.

RESULTS

LINX algorithm

The input for LINX is a base-pair-consistent segmented copy number and SV callset from the previously described tools PURPLE[8] and GRIDSS.[9] The base pair consistency means that each and every copy number change in the genome is matched precisely to an SV junction, which is represented either as a breakpoint when the partner location is known or as a single breakend when the partner location cannot be unambiguously determined.

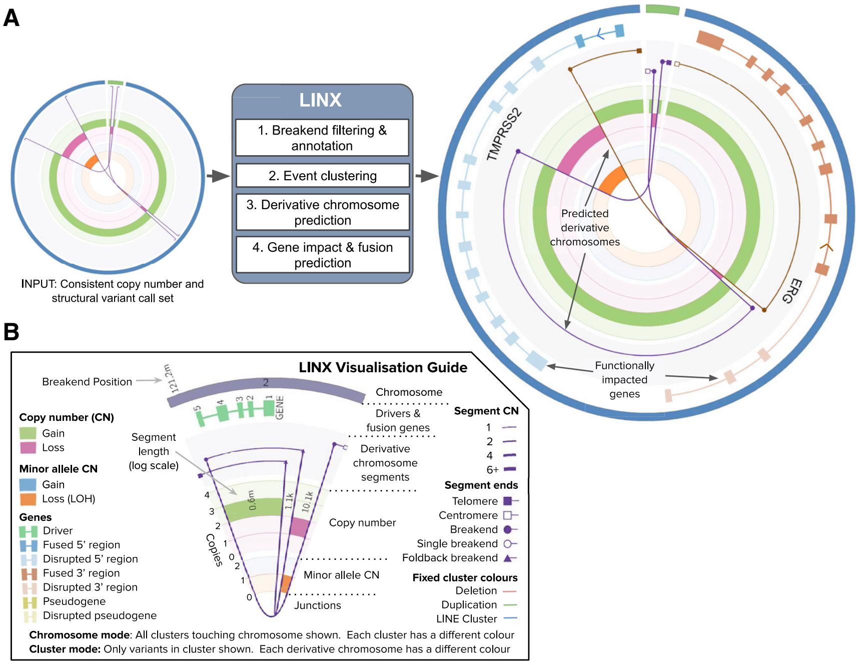


Figure 1. LINX schematic and visualizations

(A) The LINX algorithm works in four steps to annotate, cluster, chain, and determine the functional impact of an integrated copy number. The Circos on the left represents the input of LINX and shows three structural variants (purple lines) affecting two chromosomes (outer track in green and blue) with consistent copy number breakpoints (middle track showing green for gain and red for loss). The Circos on the right shows example output of LINX, including the chaining of the variants into two continuous predicted derivative chromosomes (lines in brown and purple) and a canonical TMPRSS2_ERG fusion (genes depicted in blue and light brown on second outer circle with fused exons showing darker shading) on one of the two predicted chromosomes.

(B) A detailed guide to the visualizations produced by LINX.

There are four key steps in the LINX algorithm (Figure 1; Methods S1). First, LINX annotates each breakpoint and breakend with several basic geometric and genomic properties that are important to the clustering and chaining algorithm. This includes whether each breakend is part of a foldback inversion, flanks a region of loss of heterozygosity (LOH), or is in a well-known fragile site region.[8,10] LINX also annotates well-known line element source locations[5] and identifies additional suspected mobile LINE source elements based on both the local breakpoint structure and signals of poly-A sequence insertions.

Second, LINX performs a clustering routine to group raw structural variants into distinct rearrangement "events." LINX defines a rearrangement event as one or more junctions that likely occurred proximately in time and transformed the genome from one stable configuration to another. Events can range from a simple deletion or tandem duplication to complex events, including chromothripsis or breakage fusion bridge[11] cascades. The fundamental principle for clustering in LINX is to join breakpoints where it is highly unlikely that they would have occurred independently. Rather than a single rule, such as clustering variants into events based solely on proximity[12] or variants that form a "deletion bridge,"[4] LINX employs a set of 11 independent rules in its clustering routine (Methods S1). These include clustering variants that are very close in proximity (<5 kb between breakends); clustering breakends that together delimit an LOH event, homozygous deletion, or region of high major allele copy number; clustering translocations that share common arms at both ends; clustering inversions, long deletion, and long tandem duplication variants that directly overlap each other; and clustering all foldback inversions that occur on the same chromosome arm.

Third, after resolving all variants into clusters, LINX predicts the derivative chromosome structure via a chaining algorithm. To do this, LINX considers all pairs of facing breakends on each chromosomal arm within each cluster and iteratively prioritizes which pair is most likely to be joined. The chaining logic imposes allele specific copy number constraints at all points on each chromosome and also the biological constraint that chromosomes are not permitted without a centromere unless strict criteria relating to detection of extrachromosomal DNA are met. Foldback inversions are also explicitly modeled to allow chaining of clusters of variable junction copy number and high amplification. Overall, the chaining prioritization scheme is designed to be error tolerant and aims to maximize the chance that each individual breakend is linked correctly to the next breakend on the derivative chromosome. However, due to multiple possible paths, upstream sources of error, and missing information, the prediction is representative only and, in the case of highly complex clusters, unlikely to be correct across all break junctions.

The fourth and final step in LINX is to annotate the gene impact of SV junctions to predict gene disruptions and fusions. Any breakend overlapping or in the upstream region of an Ensembl transcript[13] is annotated with its position and orientation relative to the strand of the gene and the nearest splice acceptor or donor. Gene fusions are called by searching for correctly oriented splice acceptor and donor pairs on the predicted derivative chromosome, including chained fusions that may span multiple break junctions.[14] To meet the fusion calling criteria, the breakends must also connect to viable contexts in each gene and not be terminated by further breakends in the chain on either 5′ or 3′ partner end (Methods S1). Since complex rearrangements may result in many candidate gene fusions, LINX streamlines clinical interpretation by providing a curated list of known pathogenic fusion gene pairs, as well as known promiscuous 5′ and 3′ fusion gene partners, and marks matching fusions as reportable. Finally, LINX also matches amplification, deletion, and LOH drivers called by PURPLE across a panel of well-known cancer genes (Table S1)[8] to specific SV clusters and calls additional disruption driver events in tumor suppressor genes.

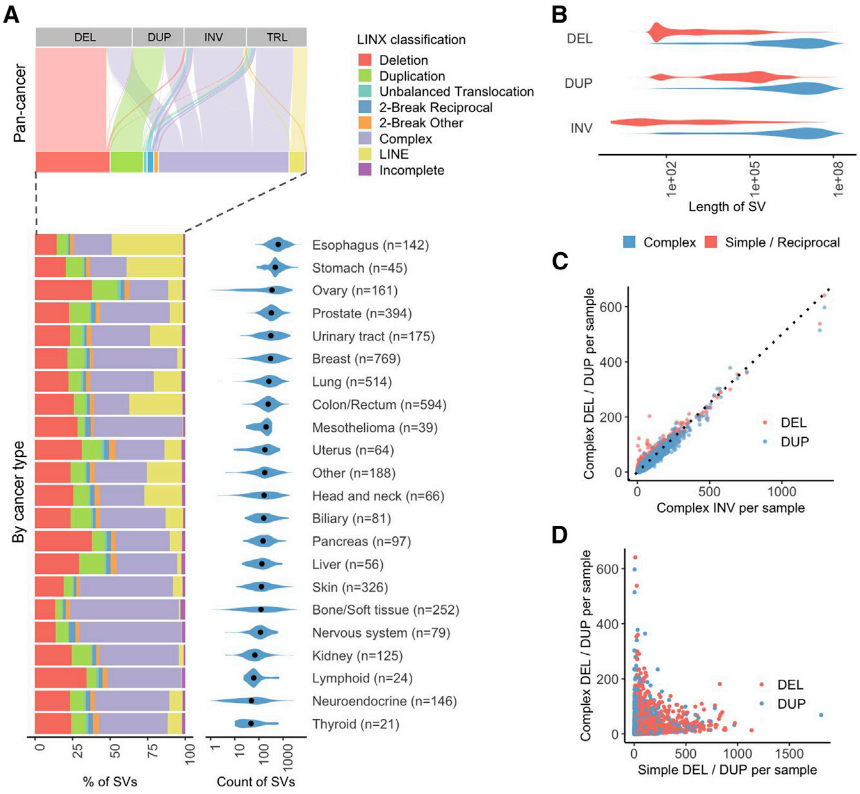


Figure 2. Landscape of genomic rearrangements

(A) Top panel shows an alluvial plot depicting the proportional assignment of each of the raw structural variant types (DEL, deletion; DUP, duplication; INV, inversion; TRL, translocation) to LINX classification. The LINX classifications are further broken down by tumor type in a relative bar chart in the left lower panel. The right lower panel shows the distribution of the number of structural variants per sample grouped by tumor type, with the black dots indicating the median values.

(B) Length distribution of notional deletion, duplications, and non-foldback inversions for both simple rearrangements and complex clusters (containing three or more variants). Note that foldback inversions have a distinct length distribution and are shown separately in Figure S4C.

(C) Counts of deletions and duplications in complex clusters per sample both closely follow a 1:2 ratio (indicated by dotted line) compared with inversions, as expected by random rearrangements following catastrophic events.

(D) Counts of simple deletions and duplications per sample are not correlated with counts of deletions and duplications in complex clusters.

See also Figure S4 and Table S3.

Pan-cancer landscape of genomic rearrangements

To demonstrate the functionality of LINX, we ran it on a pan-cancer cohort of 4,358 paired tumor-normal, whole-genome-sequenced (median of 106× and 38× paired-end sequencing coverage, respectively) adult metastatic cancer samples from Hartwig Medical Foundation (referred to as Hartwig cohort; Table S2).[8] Of these samples, 1,924 had matched whole transcriptome sequencing data, which were used for orthogonal validation where appropriate. Overall, we found a mean of 324 rearrangement junctions per sample with the highest rates in esophagus (mean = 753) and stomach (mean = 647) tumors and lowest rates in thyroid (mean = 102) and neuroendocrine (mean = 109) tumors (Figure 2A; Table S3). Event classification by LINX highlighted the diversity and tumor type specificity of rearrangement mechanisms with deletions, tandem duplications, LINE insertions, and complex events (defined as events with three or more junctions) found to be the largest classes of rearrangements in agreement with previous pan-cancer whole-genome analysis.[1] We examined each of these event classifications in detail as follows.

Classification of simple and complex rearrangement events

Classification of event types in LINX can considerably simplify interpretation of a cancer genome. An important use case is to distinguish simple events driven by a single break resulting in deletions and duplications (Figure S1) from variants that are notionally called deletions and duplications by an SV caller but may be part of a more complex event. Clean mutational profiles for simple deletions and duplications are important for downstream applications, such as signature analysis[1] and in particular homologous recombination (HR) deficiency classification,[15,16] which is associated with both short deletions and tandem duplications and may be relevant to cancer treatment.

In the Hartwig cohort, we find that lengths of deletions and duplications classified as simple events are notably shorter than those clustered in complex events (Figure 2B). Moreover, the simple deletions and duplications show distinct characteristic length peaks, which have been previously shown to be associated with *BRCA1*, *BRCA2*, and *CDK12* inactivation or *CCNE1* amplification,[17] as well as a short DUP signature that we have recently shown to be associated with colorectal tumors.[9] On the other hand, the deletions and duplications involved in complex events have length distributions closely resembling that of inversions clustered in complex events. We also find that the per-sample counts of deletions, duplications, and inversions in complex events closely follows a 1:1:2 ratio as expected from random rearrangements following a catastrophic event (Figure 2C). However, the counts of simple deletion and duplication junctions per sample were only very weakly positively correlated with those for deletions or duplications that are categorized as part of complex events (deletions r = 0.156; duplications r = 0.13; Figure 2D). Taken together, these observations suggest LINX is able to accurately distinguish between simple and complex rearrangements.

LINX annotates every cluster involving two break junctions (further referred to as two-break junction events) with a resolved type where they can be consistently chained (Figure S2) or marks as "incomplete" where they cannot form a consistent set of derivative chromosomes (Figure S3). Consistent two-break junction clusters fall into two major categories—reciprocal events (e.g., reciprocal inversions or translocations) or events with insertions of a templated sequence either in a chain or cycle.[1] We observe that two-break junction events with insertion sequences frequently involve very-short-templated sequences <1 kb in length, referred to as "genomic shards,"[18] which we find to be pervasive in cancer, constituting 14% of somatic breakpoints. Genomic shards can confound classification of otherwise simple variant types, because a short-templated insertion from another chromosome appears notionally as two translocations and can easily be misinterpreted as a reciprocal translocation or more complex event.

LINX classifies events that can be resolved as a simple deletion, tandem duplication, or translocation event with one or more inserted shards as a "synthetic" event, under the assumption that the structure is likely created by the disruption of a simple event with the insertion of the templated sequence during repair without affecting the shard donor locus. In support of this hypothesis, we find that samples with high counts of simple deletion and duplications have significantly higher ($p < 1 \times 10^{-60}$ for both) counts of synthetic deletion and duplications, respectively (Figures S4A and S4B), and furthermore, we observe the lengths of synthetic deletions and duplications to be highly consistent with the respective lengths of simple deletions and duplications (Figure S4C). Synthetic deletion and duplication events can have many different breakend topological rearrangements, depending on the source and orientation of the inserted shard (Figure S1). Insertion of genomic shards is by no means unique to simple deletion and duplication events, as we also see frequent short-templated insertion sequences in breaks of more complex events, including foldback inversion and chromothripsis events. Synthetic foldback inversions also show the same length distribution as simple foldbacks (Figure S4C).

Reciprocal events are the other major category of two-break junction events. These arise from the crossover of multiple concurrent double-stranded breaks forming either a reciprocal inversion if both breaks occur on a single chromosome (with the segment in between the two breaks repaired inverted) or a reciprocal translocation if the repair is interchromosomal. Although reciprocal inversions and translocations are found in 65% of samples in the Hartwig cohort, they are infrequent relative to other events in cancer, making up 0.8% and 0.5% of all clusters, respectively. In addition to these classical reciprocal events, we also find other configurations of reciprocal events involving two break junctions (Figure S2). One prominent configuration that we term "reciprocal duplication" involves a pair of reciprocal translocations or inversions but with breakends facing each other at both ends with substantial overlap, often multiple kilobase or even megabase in length (Figure S4D). Reciprocal duplications are significantly enriched ($p < 1 \times 10^{-60}$) in samples with strong tandem duplication signatures (Figure S4E). Furthermore, the length distribution of reciprocal duplications matches the length distribution of the signature for samples with drivers known to cause tandem duplication phenotypes, i.e., *BRCA1*, *CCNE1*, or *CDK12* drivers (Figure S4F). This suggests that these reciprocal duplication events may arise from the same process that forms tandem duplication events, likely when multiple tandem duplications occur simultaneously in a cell and, instead of repairing locally, they may cross over and create a reciprocal duplication. This observation places constraints on the mechanism by which tandem duplications may form, because it requires duplication of DNA at both loci prior to breakage and is consistent with a replication restart-bypass model,[19] but not a microhomology-mediated, break-induced replication model.[20]

Mobile element and pseudogene insertion detection

Somatic integration of LINEs is a common feature in many types of cancer, particularly esophagus and head and neck cancers.[5] A LINE insertion may involve either the transposition of a full or partial LINE source element or the transduction of a partnered or orphaned genomic region within 5 kb downstream of the LINE element. While LINE insertions are typically simple events in themselves, correct classification of these break junctions is important to accurate interpretation of the genome, as they can otherwise be mistaken as translocations and other complex events.

LINE integrations can be difficult to resolve with short read technology, because the inserted sequence is often not uniquely mappable in the genome and typically includes a Poly-A tail,[21] making assembly difficult. LINX circumvents both these issues by leveraging GRIDSS's single breakend-calling capability[9] to identify LINE insertion sites with breakend evidence for either repetitive LINE sequence, PolyA sequence, or a list of known recurrently active LINE source elements. To validate LINX's detection of mobile element insertions, we ran LINX on 75 samples from the pan-cancer analysis of whole genomes (PCAWG) pan-cancer cohort and compared LINX's LINE insertion calls with those from TraFiC-mem.[5] Overall, 339 of 564 (60%) LINX LINE insertions calls were also detected by TraFiC-mem, with TraFiC-mem calling an additional 270 insertions not found by LINX. The concordance in total LINE insertion count was very strong on a per-sample basis (Figure S5A; Table S4), with most of the private calls in both pipelines being found in the high LINE mutational burden samples (Figure S5B), suggesting that many of the private calls from both pipelines may be genuine LINE insertions.

Across the full Hartwig cohort, LINX found 76% of tumors have at least one LINE insertion event. Some tumors suffer extreme deregulation, with 6.7% of tumors having over 100 insertions and 2,241 insertions found in a single esophagus tumor sample (Figures 3A and 3B). The five most frequently inserted LINE source elements in the Hartwig cohort were all among the top six reported previously in the PCAWG pan-cancer cohort:[5] chr22:29,059,272–29,065,304, chrX:11,725,366–11,731,400, chr14:59,220,385–59,220,402, chr9:115,560,408–115,566,440, and chr6:29,920,213–29,920,223. Analysis of the precise breakend locations at these five sites reveals highly recurrent site-specific patterns of transduction (Figures 3C and S5C), where the 3′ ends of the transduced sequences are normally sourced from a handful of specific downstream sites (presumably polyadenylation sequences of alternative transcription endpoints for the LINE source element), whereas the location of 5′ side of the transduction appears to be relatively randomly distributed.

At the LINE insertion site, accurate breakpoint determination can also give insight into potential biological mechanisms. LINX finds frequent target-site duplication[22] but intriguingly finds

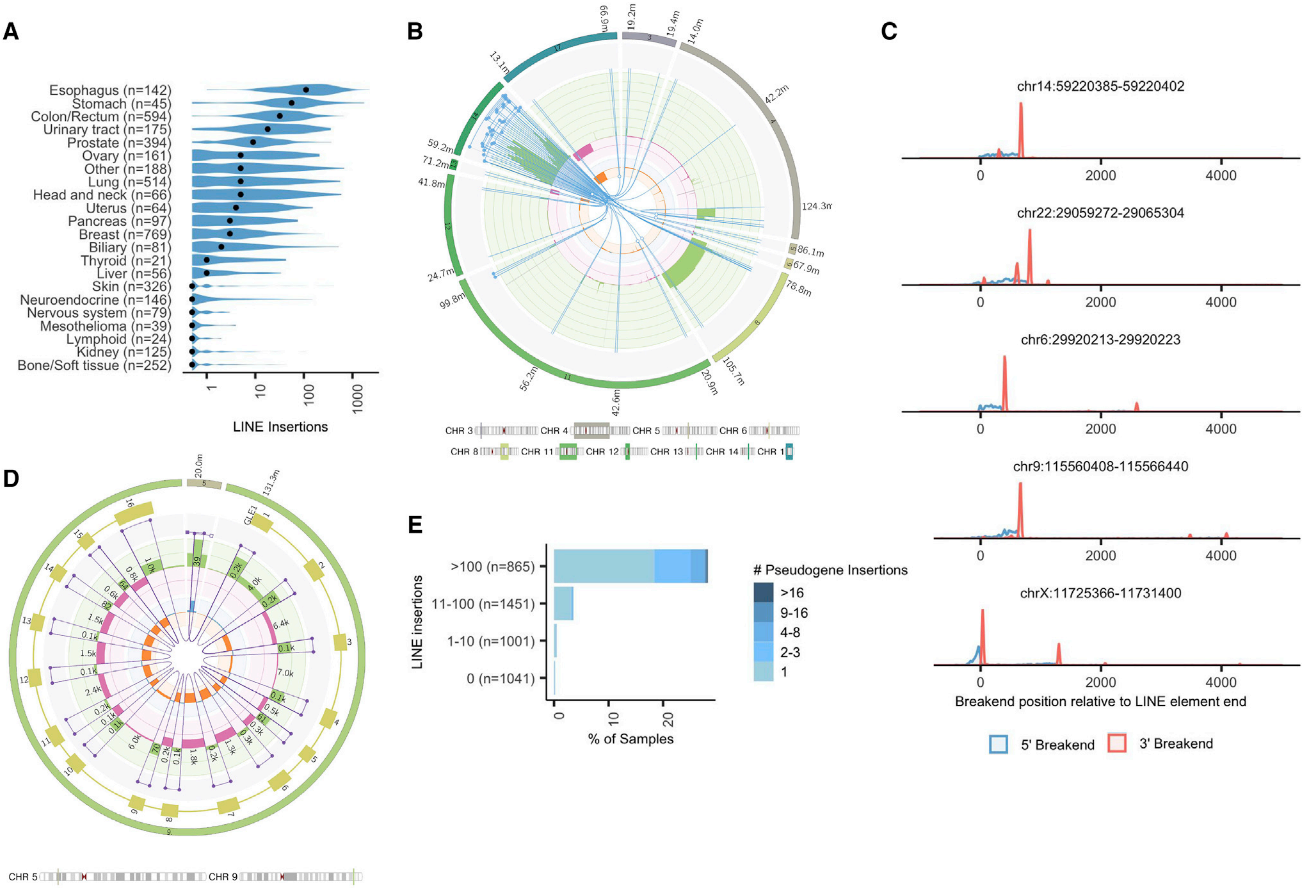


Figure 3. Mobile element insertions

(A) Violin plot showing the distribution of the number of LINE insertions per sample grouped by tumor type. Black dots indicate the median values for each tumor type.

(B) Complex LINE cluster in HMF002232B, a colorectal cancer. Overlapping segments from the LINE source element from chr14:59.2M have been inserted in at least 20 independent locations scattered throughout the genome.

(C) Histogram showing frequency of breakends positions for all mobile element transductions in Hartwig cohort originating from the five most active LINE source elements relative to the last base of the LINE source element.

(D) Pseudogene insertion of *GLE1* into an overlapping break junction on chromosome 5 in HMF002165A, a non-small cell lung cancer. All 16 exons of the *GLE1* canonical transcript are inserted, but parts of the first and last exons are lost.

(E) Samples with high numbers of LINE insertions also have high numbers of pseudogene insertions.

See also Figure S5 and Table S4.

two peaks in the distance between the insertion breakends, one at an overlap of 16 bases but also a second peak with no overlap, suggesting the possibility of two distinct breakage mechanisms for the second strand after LINE invasion (Figure S5D). Furthermore, for the 20% of insertions where LINX observed a 5′ inversion in the insertion sequence (due to twin priming),[22] only a single peak with target site duplication of 16 bases is found.

LINX can also detect somatic pseudogene insertions resulting from the activated reverse transcriptase activity associated with deregulated LINE activity in tumors.[5] LINX annotates any group of deletions that matches the exact boundaries of annotated introns as pseudogene insertions (Figure 3D). We find 577 pseudogene insertions in the Hartwig cohort, exclusively in samples with somatically activated LINE mechanisms and enriched in the samples with the most deregulated LINE activity (Figure 3E).

Complex events

LINX classifies any cluster that has three or more junctions and is not resolved as a LINE source element as "complex." Previous tools, notably ChainFinder,[4] have been developed to systematically search for complex rearrangement patterns in tumors. We compared LINX and ChainFinder across 1,479 Hartwig cohort samples and found that, while in 22% of cases, LINX and ChainFinder produced near-identical clusters, the majority of junctions clustered by LINX are left unclustered by ChainFinder, while few SVs were exclusively clustered by ChainFinder (Figures S6A and S6B). We found this to be because of two main reasons: first, ChainFinder fails to cluster a large number of junctions that are highly proximate (<5 kb between breakends; Figure S6C) and, second, LINX employs a variety of clustering techniques to link distant junctions on the same chromosome arm that are not

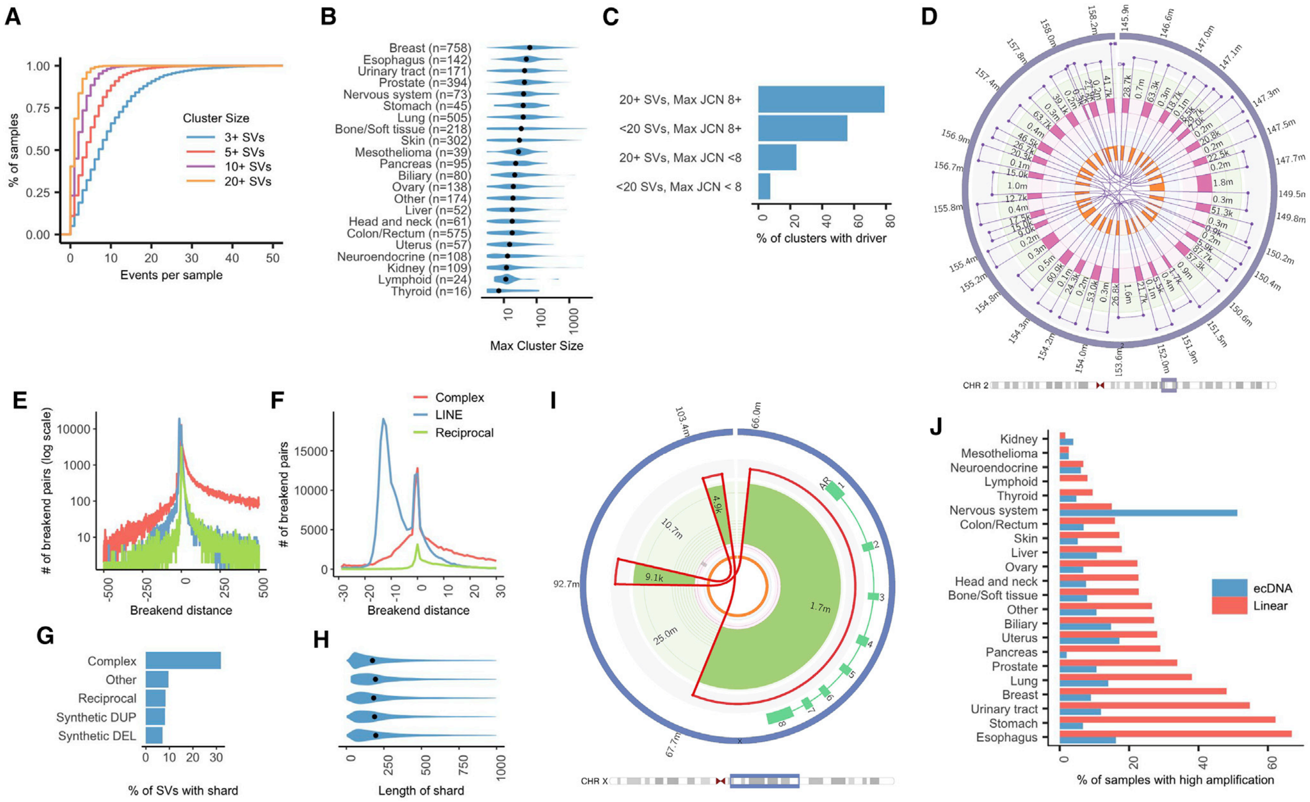


Figure 4. Complex rearrangements and high amplification

(A) Cumulative distribution function plot of count of complex rearrangement clusters per sample with at least 3, 5, 10, and 20 variants.

(B) Violin plot showing the distribution of the maximum number of variants in a single complex rearrangement cluster per sample, grouped by tumor type. Black dots indicate the median values for each tumor type.

(C) Proportion of clusters contributing to at least one amplification, deletion, homozygous disruption, or LOH driver in a panel of cancer genes by complexity of cluster and maximum JCN.

(D) Fully resolved chromothripsis event consisting of 31 structural variants affecting a 13-Mb region of chromosome 2 in HMF001571A, a prostate tumor.

(E) Counts of occurrences of *trans*-phased breakends by distance between the breakends for complex events, LINE insertions, and two-break reciprocal clusters in the range of −500 to 500 bases (log scale). Negative distances indicate overlapping breakends and duplication at the rearrangement site.

(F) Counts of occurrences of *trans*-phased breakends by distance between the breakends zoomed in to −30 to 30 bases.

(G) Proportion of variants with at least one breakend joining a shard of less than 1 kb in length by resolved type for selected resolved types.

(H) Violin plot showing the distribution of shard length by resolved type.

(I) Double minute formed by three junctions in HMF003969A, a prostate tumor, and which amplifies known oncogene, *AR*, to a copy number of approximately 23.

(J) Proportion of samples with ecDNA and linear amplifications by cancer type.

See also Figures S6 and S7 and Table S5.

captured by ChainFinder (Figure S6D). The additional variants clustered by LINX compared with ChainFinder share a strikingly similar length distribution to the variants clustered by both tools (Figure S6E), including deletions, duplications, and inversions with lengths greater than 1 Mb, which are not normally found in simple events.

Conversely, in a small proportion (1.8%) of cases, junctions are clustered by ChainFinder and not by LINX. Ninety-five percent of these are deletions and tandem duplications <1 Mb in length that may also have occurred as independent events and be inadvertently clustered by ChainFinder (Figure S6E). In line with this hypothesis, we find that 20% of the deletions clustered by ChainFinder, but not by LINX, are in known fragile sites (Figure S6F) and often are phased in *trans*, suggesting that they likely occurred in different events.[12]

Across the Hartwig cohort, we found at least one complex event in 95% of tumors and at least one event of 20 or more junctions in 60% of tumors (Figure 4A). While there are relatively few complex events in any given tumor, they account for more than half of junctions overall. Complex clusters with >100 junctions were found in all cancer types, with breast cancer having the highest median maximum complex cluster size of 62 (Figure 4B). We observe that complex events with a higher number of junctions are more likely to disrupt or amplify a putative cancer driver gene. Overall, 12.7% of all complex clusters in the cohort contributed to a LOH, amplification, deletion, or disruption driver, but this rises to 39.1% for events with 20 or more junctions and 77% for events with more than 20 junctions and high amplification (junction copy number ≥8; Figure 4C).

LINX goes further than other clustering tools in that it allows not only for complex clusters to be identified but in many of cases is able to completely resolve such events into a consistent set of derivative chromosomes, including in chains with up to 33 junctions (Figure 4D). Uniquely, and critically for accurate chaining in these complex structures, LINX utilizes phased assembly output from GRIDSS to determine whether proximate facing breakends are *cis* or *trans* phased. We observe that *trans*-phased facing breakends, causing local duplication, are common in complex events and can often extend up to several hundred bases but only rarely extend beyond 30 bases in reciprocal events and mobile insertions (Figures 4E and 4F), suggesting a fundamentally different breaking mechanism in complex events, which may cause double-stranded breaks with hundreds of bases overlap. Proximate *cis*-phased breakends are even more common than *trans*-phased and resemble in length distribution the shards detected in simple events but with much higher frequency in complex clusters (Figures 4G and 4H). We frequently observe localized regions of scarring with multiple distinct shards sourced from the same location, sometimes with overlapping template sequences.

Amplification mechanisms

Regions of high amplification are among the most complex events in tumors, as they require iterative and repeated cycles of synthesis or unequal segregation to form. There are two well-known key distinct biological mechanisms that create highly amplified rearrangements: repeated cycles of breakage fusion bridge (BFB) and stochastic amplification of circular extrachromosomal DNA by asymmetric segregation during cell division (ecDNA). ecDNA (Figures 4I and S7A) may arise from any event that creates simultaneous multiple double-stranded breaks on the same chromosomal arm, with one or more chromosomal segments repairing to form a circular structure without a centromere. BFB (Figure S7B), on the other hand, is triggered by the formation via translocation or foldback inversion of a chromosome with two centromeres, arising from either multiple concurrent double-stranded breaks or telomere erosion, and leads to duplication of chromosomal segments within a linear chromosome.

Despite these significant differences in mechanism, distinguishing between ecDNA and BFB is non-trivial based on short-read sequencing data but is essential in order to understand the diversity of amplification drivers in tumors and may be relevant to the prognosis or treatment of certain tumors.[23] The key difficulties in discrimination are that both mechanisms can leave a similar footprint, as both may arise out of complex shattering events and are highly shaped by the same selection processes, both positive (amplification of key oncogenes) and negative (constraints on amplifications of other proximate genes).

LINX employs a set of heuristics to identify subsets of clusters as likely ecDNA. The key principle used to identify ecDNA is to look for high junction copy number (JCN) structural variants adjacent to low-copy-number regions that can be chained into a closed or predominantly closed loop. LINX also checks that the high JCN cannot be explained by compounding linear amplification mechanisms, by comparing the JCN of the candidate ecDNA junctions with the maximal amplification impact of foldback inversions (hallmarks of BFB) and other junctions that link closed segments of the ecDNA to other regions of the genome (Methods S1). To validate the ecDNA heuristic, we ran LINX on a set of 13 WGS neurosphere-cultured glioblastoma samples that had been previously analyzed[24] for ecDNA with Amplicon Architect.[6] LINX and Amplicon Architect called ecDNA for an identical set of 19 oncogenes across the 13 samples (Table S5), including the 11 samples that were orthogonally validated by fluorescence *in situ* hybridization (FISH).

Applying the heuristic to the Hartwig cohort, we found ecDNA to be a relatively uncommon event present in 9.9% of all tumors, with the highest frequency in CNS tumors (51%; Figure 4J). This is lower than found in a large recent pan-cancer cohort analysis of WGS using AmpliconArchitect,[23] which found a pan-cancer prevalence of 14%. We observe that, overall, 12% of putative amplification drivers identified in the Hartwig cohort are associated with ecDNA events (Figure S7C) but that this rate increases for more highly amplified events to greater than 40% for events with maximum JCN > 32. The relative rate of ecDNA is the highest for *EGFR* (Figure S7D), but this appears to be highly specific to CNS tumors (where 87% of *EGFR* amplifications are associated with ecDNA), whereas for lung tumors (where epidermal growth factor receptor [EGFR] amplification is also common) and other cancer types, the rates of ecDNA are only 11% and 21%, respectively, similar to that of other well-known oncogenes (Figure S7E).

The high-amplification events that do not meet the ecDNA criteria are assumed to be formed via linear amplification. While we find that 76% of these events have at least one foldback inversion, suggesting a BFB process, in many events, the foldback JCN cannot explain the full amplification, and in the remaining events, LINX identifies no foldback events at all (Figure S7F). The majority of these are unlikely to be ecDNA, however, because there is no obvious set of junctions and segments that can be closed into a circle with a consistent copy number. Some events, such as the exceptionally complex amplifications of *MDM2* and *CDK4* common in liposarcoma,[3] may not fall neatly into either an ecDNA or BFB classification (Figure S7G) and have recently been proposed to be a novel rearrangement class termed ‘‘tyfonas.’’[12]

Detection of clinically relevant pathogenic rearrangements

LINX calls a diverse and comprehensive range of fusions and pathogenic rearrangements (Figures 5A and S8A–S8D). We orthogonally validated LINX’s pathogenic fusion predictions by comparing them with fusions predicted from RNA sequencing (RNA-seq) data taken from the same samples. For the RNA comparison, we used Arriba, one of the best performing RNA fusion callers,[25] using a curated list of 391 known pathogenic fusion pairs (Table S6). Across 1,924 Hartwig cohort samples with matched RNA, 148/173 in-frame fusions (86%) predicted by LINX were also found by Arriba (Figure 5B; Table S7). Of the 25 fusions not identified in RNA, 13 matched the characteristic tumor type of the known fusion pair (nine of which were *TMPRSS2-ERG* fusions in prostate cancer) and are likely to be pathogenic but with insufficient expression to be detected in the RNA. A further two cases predicted by LINX were found by Arriba but only in out-of-frame transcripts. Thirteen known pair

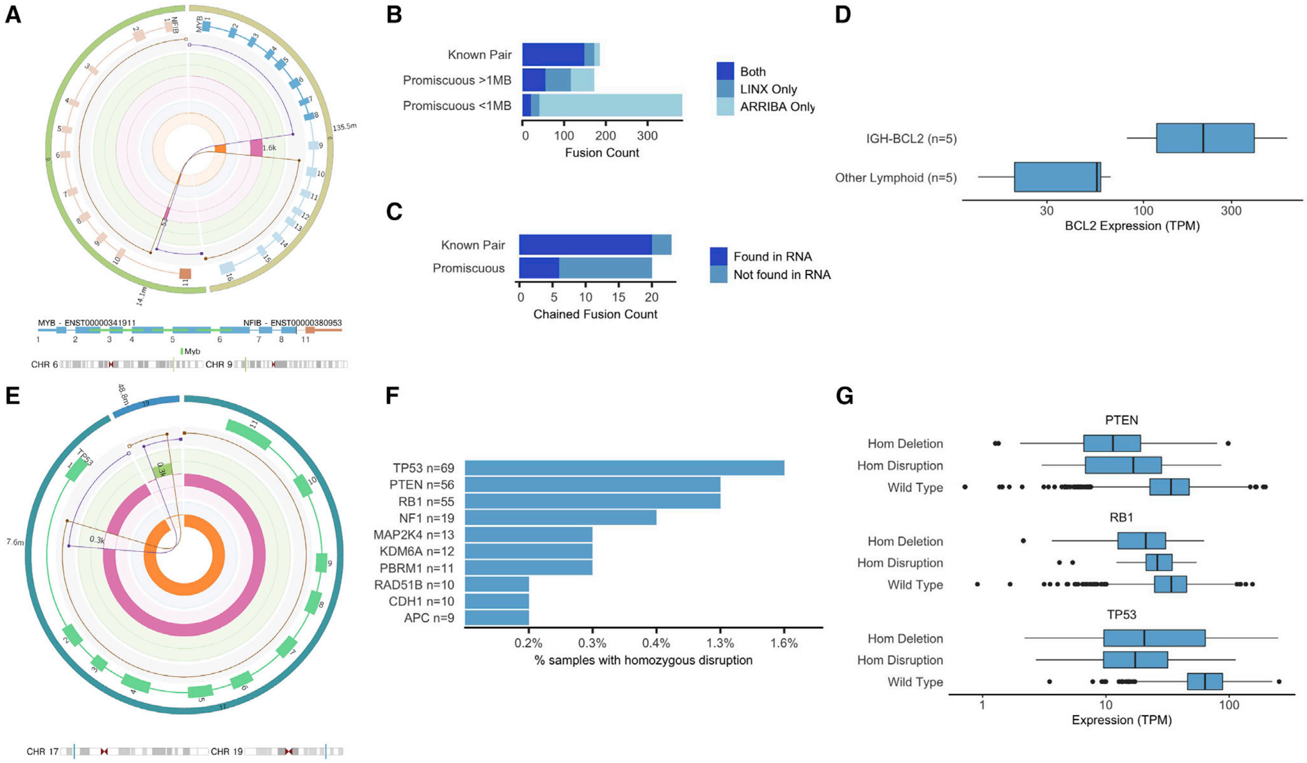


Figure 5. Clinically relevant rearrangements

(A) A *MYB-NFIB* fusion caused by a reciprocal translocation in HMF000780A, a salivary gland tumor. The translocation links exons 1–8 in *MYB* to exon 11 in *NFIB*.
(B) Comparison of LINX fusion predictions in Hartwig cohort to Arriba fusion predictions from orthogonal RNA sequencing for known pairs and promiscuous fusion partners. Promiscuous fusions of less than 1 Mb length are shown separately, as they may occur from readthrough transcripts and not be associated with a genomic rearrangement.
(C) Count of LINX chained fusion predictions for known and promiscuous fusion and whether they are also found to be expressed in RNA by Arriba.
(D) Distribution of *BCL2* expression in lymphoid samples with and without a predicted pathogenic *IGH-BCL2* rearrangement. Box: 25th–75th percentile; whiskers: data within 1.5 times the interquartile range (IQR).
(E) Reciprocal translocation affecting *TP53* in HMF001913A, a prostate tumor. The two predicted derivative chromosomes overlap by approximately 300 bases on both ends but are *trans* phased, which rules out the possibility of a templated insertion at either location. Although the *TP53* copy number alternates between one and two, no derivative chromosome contains the full gene and the gene is homozygously disrupted.
(F) Prevalence of homozygous disruption drivers for top 10 most affected tumor-suppressor genes.
(G) Distribution of gene expression in Hartwig cohort for samples with homozygous deletion, homozygous disruption, and wild type for each of *RB1*, *TP53*, and *PTEN*. box: 25th–75th percentile; whiskers: data within 1.5 times the IQR.
See also Figures S8 and S9 and Tables S6, S7, S8, and S9.

fusions were predicted by Arriba, but not by LINX, seven of which involve gene pairs less than one million bases apart on the same chromosome and may be caused by readthrough transcripts[26] or circularized RNA[27] unrelated to structural rearrangements in the DNA.

In addition to known pathogenic fusion pairs, 63 cancer-related fusion genes were curated as promiscuous 5′ and 3′ fusion partners. Among these, LINX identified a further 152 candidate in-frame fusions, 74 (49%) of which were also detected in RNA. Arriba detected 397 additional promiscuous candidates, but 86% of these were proximate on the same chromosome and are likely readthrough transcripts with no genomic rearrangement. Altogether, 43 of the 325 (13%) known and promiscuous fusion predictions were chained fusions involving multiple junctions, 26 (60%) of which were validated in the RNA-seq data (Figure 5C), highlighting the utility of chaining of derivative chromosomes for DNA fusion calling. TMPRSS2-ERG was the only fusion that LINX found to be recurrently chained in the cohort, accounting for 14 of the 43 predicted chained fusions, all in prostate cancers.

Immunoglobulin enhancer rearrangements are a distinct class of pathogenic rearrangements, common in B cell tumors where errors in VDJ recombination and/or isoform switching in the *IGH*, *IGK*, and *IGL* regions may lead to pathogenic rearrangements, driving high expression of known oncogenes through regulatory element repositioning.[28] Although these typically do not make a novel protein fusion product, LINX predicts these pathogenic rearrangements based on the breakend in the *IGH*, *IGK*, and *IGL* regions with orientation and position matching locations commonly observed in B cell tumors.[28] Among 10 lymphoid samples with matching RNA in the cohort, LINX found six such rearrangements, including five cases of *IGH-BCL2* and one case of *IGH-MYC*. The five identified samples with *IGH-BCL2* rearrangements have significantly higher expression (p = 0.008) of *BCL2* than the five lymphoid samples with no *BCL2* rearrangement detected (Figure 5D).

LINX also identifies disruptive intragenic rearrangements that may cause exonic deletions and duplications. Our knowledge base includes nine such rearrangements known to be pathogenic and two that we have deemed likely pathogenic due to high recurrence in the Hartwig cohort. Three of the known pathogenic exon rearrangements were detected by LINX in at least five samples with paired RNA in our cohort: *EGFRvII* (n = 6), *EGFRvIII* (n = 14), and *CTNNB1* exon three deletion (n = 6). In all cases with an event detected by LINX in the DNA, we found RNA fragments that supported novel splice junctions in the matched RNA (Figure S8E). Only one other sample in the complete cohort (n = 1,924) had more than one fragment supporting any of these alternative splice junctions (a gastrointestinal stromal tumor with three fragments supporting EGFRvII but with no evidence of rearrangement in EGFR), suggesting a low false-negative rate in LINX.

In addition to producing novel oncogenic proteins and overexpression of well-known oncogenes, rearrangements may also lead to tumorigenesis by disrupting the function of tumor-suppressor genes. To capture this, LINX annotates every breakend that overlaps a gene, determines whether it is disruptive to the gene, and reports the number of undisrupted copies. In cases of reciprocal translocations (Figure 5E), reciprocal inversions (Figure S9A), complex break events (Figure S9B), or tandem duplications that overlap at least one exon (Figure S9C), a gene may be disrupted on all remaining copies, even though the copy number is greater than zero for all exonic bases.[29] We term this type of genomic rearrangement a "homozygous disruption." Homozygous disruptions cannot readily be detected by standard panel or whole-exome sequencing, since intronic sequences are typically not included in such panels and they are copy neutral in exonic regions.

We find homozygous disruptions to be a common driver in the Hartwig cohort, with 9% of samples containing at least one homozygous disruption in a panel of 448 curated cancer-related genes (Table S8). Three well-known tumor-suppressor genes had homozygous disruptions in more than 1% of the cohort: *TP53* (n = 69), *PTEN* (n = 56), and *RB1* (n = 55; Figure 5F). Supporting the functional impact of these events, we found significantly lower expression for each of these genes (TP53: $p = 2 \times 10^{-16}$; PTEN: $p = 2 \times 10^{-6}$; RB1: $p = 2 \times 10^{-3}$) in samples with predicted homozygous disruptions compared with samples with at least one intact copy (Figure 5G) and similar mean fold change in expression compared with samples with homozygous deletions (TP53: 0.30 versus 0.40; PTEN: 0.47 versus 0.37; RB1: 0.68 versus 0.60 for disruptions and deletions, respectively). We also performed a genome-wide search for genes with enrichment of homozygous disruptions and found 35 significantly enriched genes, including 16 well-known tumor suppressors, 14 genes immediately adjacent to tumor-suppressor genes, and three highly recurrent oncogenic fusion partners (Table S9). Intriguingly, we found an additional two genes also enriched in homozygous disruptions, but not widely characterized as tumor-suppressor genes: *PSIP1* (five observations; q = 0.006), which has previously also shown to be significantly enriched in truncating point mutations,[30] and *USP43* (six observations; q = 0.01), a recently proposed tumor suppressor.[31]

Visualization

LINX produces detailed visualizations of the rearrangements in the tumor genome that allow further insights into complex rearrangements. LINX supports either drawing all rearrangements in a cluster or all the rearrangements on a chromosome, creating an integrated Circos output[32] showing copy number changes, clustered SVs, the derivative chromosome predictions, and impacted genes, including protein domain annotation for gene fusions, all on the same diagram. The visualizations use a log-based position scaling between events so that small- and large-scale structures can both be inspected on a single chart. Combined with the circular representation, these features allow unprecedented resolution of complex structures across a broad array of event types, including chromoplexy (Figure 6A) and complex BFB amplification events (Figure 6B). Methods S1 includes a walkthrough and explanation of all LINX figures, covering the complete SV landscape of the COLO829T melanoma cancer cell line, which has been proposed as a somatic reference standard for cancer-genome sequencing.[33,34]

Evaluation on an independent cohort. To assess broader utility of the tool set and the reproducibility of our results, we compared the findings on the Hartwig cohort with a subset of 1,541 samples from the independently sequenced PCAWG pan-cancer cohort (Table S10).[35] The PCAWG samples analyzed also cover a diverse range of tumor types but, unlike the Hartwig cohort, contain almost exclusively primary cancer samples and are sequenced to a lower average coverage of depth (38×–60× PCAWG compared with median 106× for HMF).

We observed largely the same structural variant patterns across the two cohorts (Figure S10A). The length distributions of deletions, duplications, and inversions were highly similar for both simple and complex events across the two cohorts (Figure S10B). We also observed a very similar preponderance and length distribution of genomic shards across all event types (Figure S10C). Furthermore, we found that the length distributions of the synthetic events in the PCAWG cohort closely replicated the results found in the Hartwig cohort (Figure S10D). Likewise, the reciprocal duplication events we identified in the Hartwig cohort were also present in PCAWG, with the same length patterns of tandem duplication signatures for samples with *BRCA1*, *CDK12*, and *CCNE1* drivers (Figure S10E). Driver-related rearrangement patterns were also similar between the PCAWG and Hartwig cohorts. While the overall rates of samples with high-amplification events were lower in the primary cancers (22% PCAWG; 41% HMF), the proportion accounted for by ecDNA was similar (28% PCAWG; 24% HMF; Figure S10F). We also found homozygous disruptions events impacting tumor-suppressor genes (TSGs) in the PCAWG cohort. Indeed, the top four driver genes with putative homozygous disruption drivers were the same in both datasets (Figure S10G).

Overall, the high reproducibility of these results in the independently sequenced PCAWG cohort lends weight both to the utility of LINX and the universality of the observed patterns across both metastatic and primary cancers.

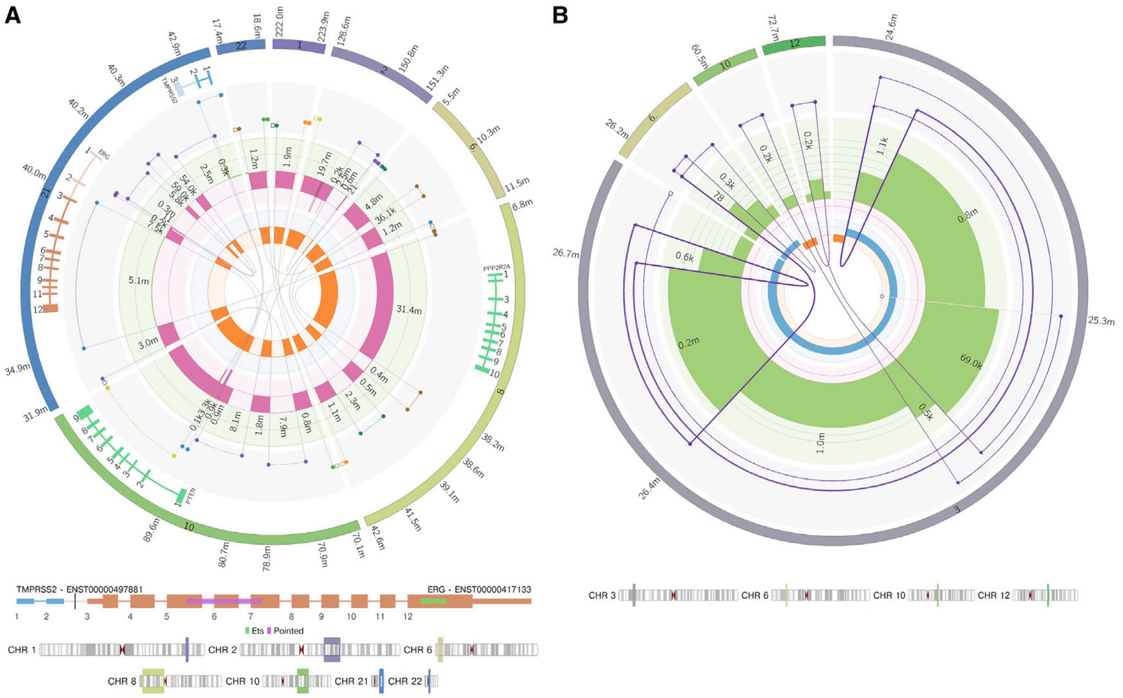


Figure 6. Complex event visualization

(A) Chromoplexy-like cluster formed from 19 break junctions across seven chromosomes in HMF001596B, a prostate tumor. The rearrangement leads to three distinct putative drivers in a single event, including a chained *TMPRSS2-ERG* fusion with two hops; a loss of heterozygosity for *PPP2R2A*, which also has a stop-gained point mutation (not shown); and an intronic homozygous disruption of *PTEN*.

(B) Breakage fusion bridge event affecting the P arm of chromosome 3 in the melanoma cell line COLO829T. The predicted derivative chromosome has a copy number of two and can be traced outwards starting from the centromere on chromosome 3, traversing two simple foldbacks and two chained foldbacks and finishing on a single breakend at chr3:25.3M, which from the insert sequence can be inferred to be connected to a centromeric satellite region (likely chromosome 1, which has a copy number gain of two over the centromere from P to Q arm and which appears to be connected to chromosome 3 in unpublished SKY karyotype figures; http://www.pawefish.path.cam.ac.uk/OtherCellLineDescriptions/COLO829.html). One chained foldback at chr3:26.4M includes a genomic shard from chr6 of approximately 400 bases, which has itself been replicated and internally disrupted by the foldback event. The other chained foldback at chr3:25.4M includes two consecutive genomic shards inserted from chromosome 10 and 12 of approximately 200 bases each.

DISCUSSION

We have shown that LINX can help understand highly rearranged cancer genomes in multiple ways. Other recent publications on complex somatic rearrangements[1,12] have developed tools, such as ClusterSV and JaBbA, that have significant feature overlap with LINX. Each of these approaches also utilizes a base-pair-consistent SV/CN call set to cluster SVs, classify certain types of rearrangements, and assess downstream impact.

However, LINX differs from existing approaches in several key aspects. First, LINX clusters use copy number consistency constraints in addition to SV proximity. Second, LINX chains SVs to reconstruct the derivative chromosomes caused by each rearrangement event, including partial reconstruction for incomplete events. Third, LINX performs comprehensive classification. Every SV is classified, including mobile element translocations. Fourth, LINX utilizes single-breakend SV calls. The single-breakend repeat annotations provided by GRIDSS enable LINX to classify mobile element translocations as well as cluster complex events overlapping centromeric repeats. Fifth, LINX's rearrangement model allows for genomic shards to be inserted in any event type. The size distribution of sharded events indicates this approach is sound, at least for simple events, and this approach considerably simplifies the classification scheme. Sixth, LINX utilizes a nonlinear Circos-style visualization format that enables even quite complex rearrangements to be visually interpretable. Finally, LINX provides the most comprehensive genomic rearrangement functional impact analysis currently available. To the best of our knowledge, LINX is the only tool that reports homozygous disruptions and the only tool that can identify chained fusions from DNA-seq data alone, both of which can lead to clinically relevant rearrangements in tumors.

The challenges in understanding the complexity of rearrangements in tumor genomes can be daunting. The diversity of overlapping or converging biological mechanisms that may cause similar rearrangement patterns means that it may be perilous to analyze any one rearrangement as a standalone analysis. By exhaustively classifying all rearrangements, LINX is a robust foundation for more detailed analysis of specific rearrangement patterns, including structural variant signatures, complex shattering events, and high-amplification drivers as well as dissection of the underlying molecular mechanisms, DNA replication, and repair components involved. The full LINX analysis results on the Hartwig cohort are available via data request and can be paired with clinical data and other whole-genome analyses for further in-depth research.

WGS offers the promise of a single comprehensive test for all genomic alterations for both routine diagnostics and future biomarker discovery. LINX takes a step toward that goal by both comprehensively calling clinically relevant fusions from DNA with similar precision and sensitivity to gold standard RNA-seq methods and by identifying homozygous disruptions, an important class of drivers of tumorigenesis that cannot readily be detected by standard-of-care methods.

Limitations of the study

There are many potential sources of error that can confound correct interpretation of complex genomic rearrangements, including sample preparation, sequencing errors and coverage biases (such as GC bias), inaccurate fitting of sample purity and ploidy, false-positive or false-negative structural variant calls, and inaccurate local copy number measurement. Depth of coverage and sequencing quality are important considerations here. While we have shown that LINX can find highly similar results on the PCAWG dataset, which has, on average, half the sequencing coverage of the Hartwig cohort, lower depth coverage and/or lower quality sequencing is associated with higher false-negative rates of structural variants[9] and will result in less complete reconstructions.

Furthermore, while LINX has been optimized for short-read technology, the short-read length is ultimately the key limitation in interpretation, because it limits the phasing of proximate variants and accurate identification of events in long repetitive regions. Nevertheless, in practice, LINX is able to resolve many structures via various chaining and clustering heuristics, but for more complex events, particularly highly rearranged focal regions, errors are inevitable and the chaining is only partial and representative. While we have performed extensive comparison of LINX against other tools and can validate some of LINX's chaining predictions orthogonally via RNA evidence for chained fusions, there are, as yet, no representative tumor genomes with a fully resolved chromosomal structure for comparison as a truth set. Long-read sequencing technologies[36] can phase more distant breakpoints and are likely better suited for resolving complex events, although those technologies typically perform less well for small variant detection. Pairing short- and long-read technologies will no doubt lead to further advances in our understanding of the mechanisms and role of genomic rearrangements in tumorigenesis.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2022.100112.

ACKNOWLEDGMENTS

This publication and the underlying study have been made possible partly on the basis of the data that Hartwig Medical Foundation and the Center of Personalised Cancer Treatment (CPCT) have made available to the study. Whole-genome sequencing data were analyzed from glioblastoma neurosphere lines, which were established at Henry Ford Healthcare System. Funding was provided from Hartwig Medical Foundation.

AUTHOR CONTRIBUTIONS

Conceptualization, resources, methodology, data curation, and formal analysis, P.P., C.S., and D.L.C.; software and visualization, C.S., J.B., and D.L.C.; supervision, A.T.P., E.C., and M.J.C.; validation, M.W.; writing – original draft, P.P.; writing – review & editing, all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 13, 2021
Revised: September 29, 2021
Accepted: February 25, 2022
Published: March 22, 2022

REFERENCES

1. Li, Y., Roberts, N.D., Wala, J.A., Shapira, O., Schumacher, S.E., Kumar, K., Khurana, E., Waszak, S., Korbel, J.O., Haber, J.E., et al. (2020). Patterns of somatic structural variation in human cancer genomes. Nature *578*, 112–121.
2. Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A., et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell *144*, 27–40.
3. Garsed, D.W., Marshall, O.J., Corbin, V.D.A., Hsu, A., Di Stefano, L., Schröder, J., Li, J., Feng, Z.-P., Kim, B.W., Kowarsky, M., et al. (2014). The architecture and evolution of cancer neochromosomes. Cancer Cell *26*, 653–667.
4. Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M., et al. (2013). Punctuated evolution of prostate cancer genomes. Cell *153*, 666–677.
5. Rodriguez-Martin, B., Alvarez, E.G., Baez-Ortega, A., Zamora, J., Supek, F., Demeulemeester, J., Santamarina, M., Ju, Y.S., Temes, J., Garcia-Souto, D., et al. (2020). Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nat. Genet. *52*, 306–319.

6. Deshpande, V., Luebeck, J., Nguyen, N.-P.D., Bakhtiari, M., Turner, K.M., Schwab, R., Carter, H., Mischel, P.S., and Bafna, V. (2019). Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. Nat. Commun. *10*, 392.

7. Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47–54.

8. Priestley, P., Baber, J., Lolkema, M.P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., et al. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. Nature *575*, 210–216.

9. Cameron, D.L., Baber, J., Shale, C., Valle-Inclan, J.E., Besselink, N., van Hoeck, A., Janssen, R., Cuppen, E., Priestley, P., and Papenfuss, A.T. (2021). GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. Genome Biol. *22*, 202.

10. Dillon, L.W., Burrow, A.A., and Wang, Y.-H. (2010). DNA instability at chromosomal fragile sites in cancer. Curr. Genomics *11*, 326–337.

11. McClintock, B. (1941). The stability of broken ends of chromosomes in Zea mays. Genetics *26*, 234–282.

12. Hadi, K., Yao, X., Behr, J.M., Deshpande, A., Xanthopoulakis, C., Tian, H., Kudman, S., Rosiene, J., Darmofal, M., DeRose, J., et al. (2020). Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. Cell *183*, 197–210.e32.

13. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., et al. (2020). Ensembl 2020. Nucleic Acids Res. *48*, D682–D688.

14. Anderson, N.D., de Borja, R., Young, M.D., Fuligni, F., Rosic, A., Roberts, N.D., Hajjar, S., Layeghifard, M., Novokmet, A., Kowalski, P.E., et al. (2018). Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors. Science *361*, eaam8419.

15. Davies, H., Glodzik, D., Morganella, S., Yates, L.R., Staaf, J., Zou, X., Ramakrishna, M., Martin, S., Boyault, S., Sieuwerts, A.M., et al. (2017). HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat. Med. *23*, 517–525.

16. Nguyen, L., Martens, J.W.M., Van Hoeck, A., and Cuppen, E. (2020). Pan-cancer landscape of homologous recombination deficiency. Nat. Commun. *11*, 5584.

17. Menghi, F., Barthel, F.P., Yadav, V., Tang, M., Ji, B., Tang, Z., Carter, G.W., Ruan, Y., Scully, R., Verhaak, R.G.W., et al. (2018). The tandem duplicator phenotype is a prevalent genome-wide cancer configuration driven by distinct gene mutations. Cancer Cell *34*, 197–210.e5.

18. Bignell, G.R., Santarius, T., Pole, J.C.M., Butler, A.P., Perry, J., Pleasance, E., Greenman, C., Menzies, A., Taylor, S., Edkins, S., et al. (2007). Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. Genome Res. *17*, 1296–1303.

19. Willis, N.A., Frock, R.L., Menghi, F., Duffey, E.E., Panday, A., Camacho, V., Hasty, E.P., Liu, E.T., Alt, F.W., and Scully, R. (2017). Mechanism of tandem duplication formation in BRCA1-mutant cells. Nature *551*, 590–595.

20. Hastings, P.J., Ira, G., and Lupski, J.R. (2009). A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet. *5*, e1000327.

21. Tubio, J.M.C., Li, Y., Ju, Y.S., Martincorena, I., Cooke, S.L., Tojo, M., Gundem, G., Pipinikas, C.P., Zamora, J., Raine, K., et al. (2014). Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science *345*, 1251343.

22. Ostertag, E.M., and Kazazian, H.H., Jr. (2001). Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res. *11*, 2059–2065.

23. Kim, H., Nguyen, N.-P., Turner, K., Wu, S., Gujar, A.D., Luebeck, J., Liu, J., Deshpande, V., Rajkumar, U., Namburi, S., et al. (2020). Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. Nat. Genet. *52*, 891–897.

24. deCarvalho, A.C., Kim, H., Poisson, L.M., Winn, M.E., Mueller, C., Cherba, D., Koeman, J., Seth, S., Protopopov, A., Felicella, M., et al. (2018). Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. Nat. Genet. *50*, 708–717.

25. Haas, B.J., Dobin, A., Li, B., Stransky, N., Pochet, N., and Regev, A. (2019). Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol. *20*, 213.

26. He, Y., Yuan, C., Chen, L., Lei, M., Zellmer, L., Huang, H., and Liao, D.J. (2018). Transcriptional-readthrough RNAs reflect the phenomenon of "A gene contains gene(s)" or "gene(s) within a gene" in the human genome, and thus are not chimeric RNAs. Genes *9*, 40.

27. Yu, C.-Y., and Kuo, H.-C. (2019). The emerging roles and functions of circular RNAs and their generation. J. Biomed. Sci. *26*, 29.

28. Chong, L.C., Ben-Neriah, S., Slack, G.W., Freeman, C., Ennishi, D., Mottok, A., Collinge, B., Abrisqueta, P., Farinha, P., Boyle, M., et al. (2018). High-resolution architecture and partner genes of MYC rearrangements in lymphoma with DLBCL morphology. Blood Adv. *2*, 2755–2765.

29. Patch, A.-M., Christie, E.L., Etemadmoghadam, D., Garsed, D.W., George, J., Fereday, S., Nones, K., Cowin, P., Alsop, K., Bailey, P.J., et al. (2015). Whole-genome characterization of chemoresistant ovarian cancer. Nature *521*, 489–494.

30. Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2018). Universal patterns of selection in cancer and somatic tissues. Cell *173*, 1823.

31. He, L., Liu, X., Yang, J., Li, W., Liu, S., Liu, X., Yang, Z., Ren, J., Wang, Y., Shan, L., et al. (2018). Imbalance of the reciprocally inhibitory loop between the ubiquitin-specific protease USP43 and EGFR/PI3K/AKT drives breast carcinogenesis. Cell Res. *28*, 934–951.

32. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

33. Craig, D.W., Nasser, S., Corbett, R., Chan, S.K., Murray, L., Legendre, C., Tembe, W., Adkins, J., Kim, N., Wong, S., et al. (2016). A somatic reference standard for cancer genome sequencing. Sci. Rep. *6*, 1–11.

34. Valle-Inclan J.E., Besselink N.J.M., de Bruijn E., Cameron D.L., Ebler J., Kutzera J., et al. A multi-platform reference for somatic structural variation detection. Preprint at bioRxiv. https://doi.org/10.1101/2020.10.15.340497.

35. Consortium, T.I.P.-C.A. of W.G., and the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93.

36. Sedlazeck, F.J., Lee, H., Darby, C.A., and Schatz, M.C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat. Rev. Genet. *19*, 329–346.

37. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
HMF WGS/WTS BAMs	Priestley et al., 2019[8]	https://www.hartwigmedicalfoundation.nl/en/database/
PCAWG WGS BAMs	Consortium and The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020[35]	http://dcc.icgc.org/pcawg/
WGS glioblastoma neurosphere cultures BAMs	deCarvalho et al., 2018[24]	EGA accession: EGAS00001001878
Software and algorithms		
LINX v1.12	This paper	https://github.com/hartwigmedical/hmftools/tree/master/linx
GRIDSS2 v2.9.3	Cameron et al., 2021[9]	https://github.com/PapenfussLab/gridss
PURPLE v2.48	Priestley et al., 2019[8]	https://github.com/hartwigmedical/hmftools/tree/master/purple
STAR 2.7.3a	Dobin et al., 2013[37]	https://github.com/alexdobin/STAR
Isofox v1.0	Hartwig Medical Foundation	https://github.com/hartwigmedical/hmftools/tree/master/isofox
ChainFinder v1.0.1	Baca et al., 2013[4]	https://software.broadinstitute.org/cancer/cga/chainfinder
Circos	Krzywinski et al. 2009[32]	http://circos.ca/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Peter Priestley (p.priestley@hartwigmedicalfoundation.nl).

Materials availability

This study did not generate any new reagents.

Data and code availability

All raw (BAM), analysed (VCF, SV, purity copy number data) germline, and somatic genomic data and LINX results from the Hartwig cohort were obtained from the Hartwig Medical Foundation (Data request DR-005). Standardized procedures and request forms for access to this data, including LINX analysis results, can be found at https://www.hartwigmedicalfoundation.nl/en.

LINX is freely available as open source software from the Hartwig Medical Foundation (https://github.com/hartwigmedical/hmftools/tree/master/linx) under a GPLv3 license. Reference data required to run LINX on hg19 or hg38 is available from https://resources.hartwigmedicalfoundation.nl. LINX can be run from raw paired tumor-normal FASTQ files as part of Hartwig's open source cloud-based cancer analysis pipeline (https://github.com/hartwigmedical/platinum). Alternatively, a docker image is available from dockerhub as gridss/gridss-purple-linx to run GRIDSS, PURPLE, and LINX together from tumor and normal BAMs.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The patient cohort was derived from the Hartwig Medical Foundation Cohort for which the sample collection and whole genome sequencing and alignment to the GRCH37 reference genome has previously been described.[8] We filtered for the highest purity sample from each patient from tumor samples with purity $\geq$ 20% and with no QC warnings or failures, yielding 4,378 paired tumor-normal whole genome samples in total. An additional 1,774 paired tumor-normal sample BAMS were obtained from PCAWG, 1,541 of which passed QC warnings and purity filters. For 1,924 HMF samples paired whole transcriptome sequence data were also analyzed.

METHOD DETAILS

Analysis of structural variation and copy number alterations

GRIDSS[9] v2.93 and PURPLE[8] v2.48 were used for copy number and structural variant inputs for LINX. LINX v1.12 was used for all analyses in this paper and is described in detail in Methods S1.

RNA validation

The RNA-seq was aligned to the GRCH37 genome using STAR 2.7.3a.[37] Gene expression was calculated using Isofox v1.0, which uses an expectation maximisation algorithm to estimate transcript abundance from genome aligned RNA-seq data, with default parameters. Isofox was also used to count the RNA fragments supporting novel splice junctions predicted in LINX for exon deletions and duplications. Isofox is described in detail at https://github.com/hartwigmedical/hmftools/tree/master/isofox.

Known pathogenic pair and promiscuous gene fusions predictions in the DNA were compared to passing fusion calls in the RNA by Arriba (https://github.com/suhrig/arriba). Fusions were considered to be matched if the gene pair matched between RNA and DNA. Mean TPM fold change was calculated as 2 to the power of the difference in mean(log2(TPM)) between groups of samples.

Complex event validation

We compared LINX to ChainFinder[4] v1.0.1 on 2,840 samples from the Hartwig cohort. ChainFinder was run with default parameters. Both LINX and ChainFinder were run using the same GRIDSS/PURPLE input data. Only 1,479 samples for which ChainFinder completed within 24 hours were included in the comparison. ChainFinder clusters of 3 or more variants were considered equivalent to LINX's COMPLEX classification. For each individual variant we determined whether it was clustered in LINX, in ChainFinder or in both as well as the size of the cluster in each tool.

LINE insertion validation

We ran LINX on 75 WGS samples from the PCAWG cohort (Table S2) which had previously been run with TraFiC-mem.[5] Insertions were considered matched between the tools if the predicted insertion site was within 50 bases.

ecDNA validation

We ran LINX on 13 previously analysed[24] WGS glioblastoma neurosphere cultures sequenced to ~10x depth and compared the ecDNA predictions of Linx to those of the AmpliconArchitect tool and FISH. We matched the ecDNA predictions by amplified oncogene per sample.

Genes enriched in homozygous disruptions

We estimated a background rate of homozygous disruptions by dividing the total number of observed homozygous disruptions across the full Hartwig cohort by the total length of all annotated genes in the Hartwig cohort. For each gene we then compared the observed number of homozygous distribution to the expected number taking into account the global rate and the length of the specific genes using a Poisson distribution and correcting for false discovery. Genes with a false discovery rate of less than 0.1 were reported.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical tests are described in figure legends. All significance values presented for comparisons of both gene expression and counts of rearrangement types are calculated using a two-tailed Mann–Whitney U-test.

Cell Genomics, Volume *2*

Supplemental information

Unscrambling cancer genomes via integrated analysis of structural variation and copy number

Charles Shale, Daniel L. Cameron, Jonathan Baber, Marie Wong, Mark J. Cowley, Anthony T. Papenfuss, Edwin Cuppen, and Peter Priestley

Supplemental information

Unscrambling cancer genomes via integrated analysis of structural variation and copy number

Charles Shale, Daniel L. Cameron, Jonathan Baber, Marie Wong, Mark J. Cowley, Anthony T. Papenfuss, Edwin Cuppen, Peter Priestley

Contents

Supplementary figures

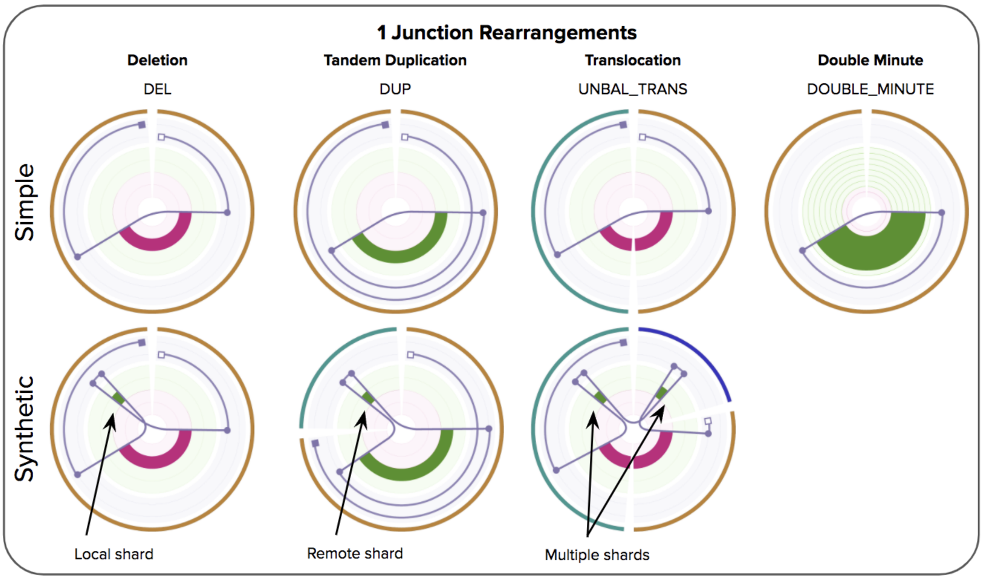


FIGURE S1: Consistent single junction rearrangements, related to STAR Methods.
A catalog of consistent 1 junction structures in LINX. The upper row shows the 4 consistent chromosome structures that can be formed from a single junction including a simple deletion, tandem duplication, unbalanced translocation or single junction double minute . The lower 3 images show examples of equivalent multiple break junction events that can be treated as equivalent synthetic versions of the above 1 junction events with one or more genomic shards (<1000 base segments of templated DNA) inserted.

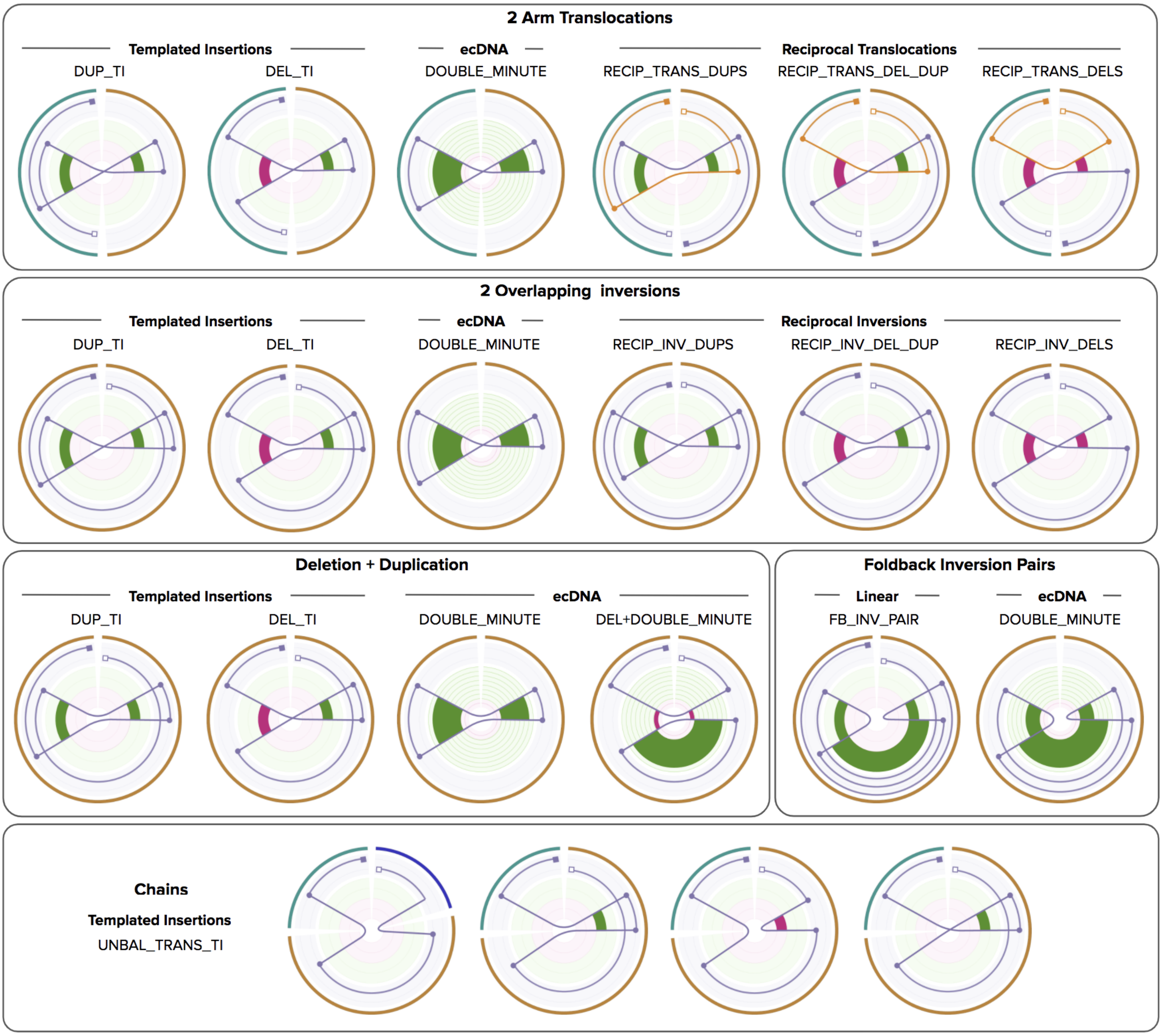


FIGURE S2: Consistent 2 junction rearrangements, related to STAR Methods.
A catalog of consistent 2 junction rearrangements in LINX. The structures are organised by the geometry of the break junctions with the resultant derivative chromosomal structure dependent on both the orientation of the breakends and whether they are phased. Pairs of common arm translocations (top panel) and overlapping inversions (2nd panel) can each form 3 different topologies of reciprocal events, 2 topologies of templated insertions and 1 double minute topology. Overlapping deletions and duplications may form templated insertion and double minutes, but not reciprocal structures (third panel, left side). Mixed combinations of translocations and local junctions (bottom panel) create a chained translocation with templated insertion (classified as 2-break other in LINX), Non overlapping foldback inversions (third panel, right side) may form either a double minute or other linear amplification structure.

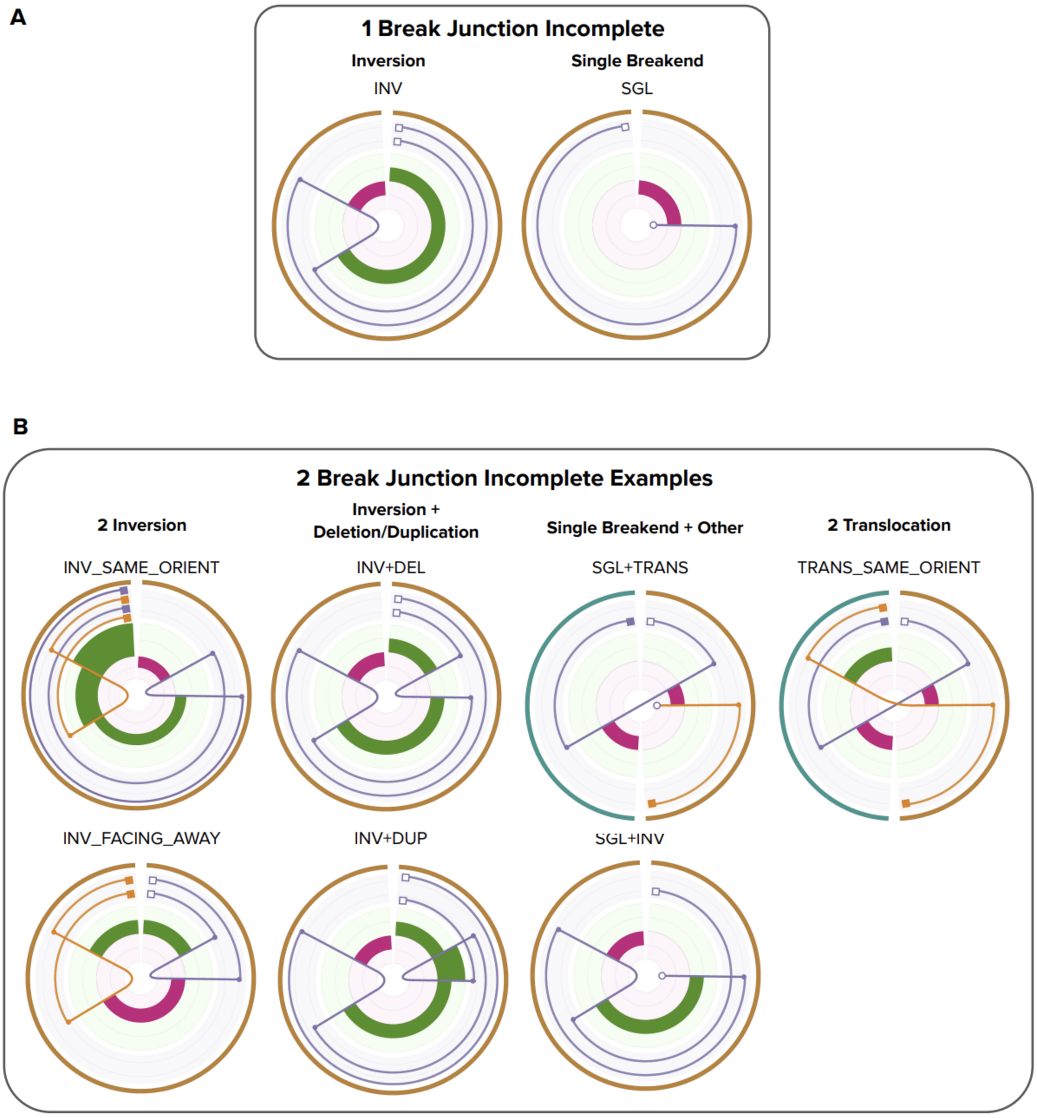


FIGURE S3: Incomplete 1 and 2 junction rearrangements, related to STAR Methods
Examples of incomplete 1 and 2 junction rearrangements in LINX. A cluster is considered incomplete in LINX if it does not create a set of 1 or more consistent derivative chromosomes that each link a telomere and centromere. A rearrangement event, such as an inversion linking 2 centromeres or 2 telomeres (without an intervening centromere) cannot guarantee equal division at mitosis and hence is likely to be inherently unstable.

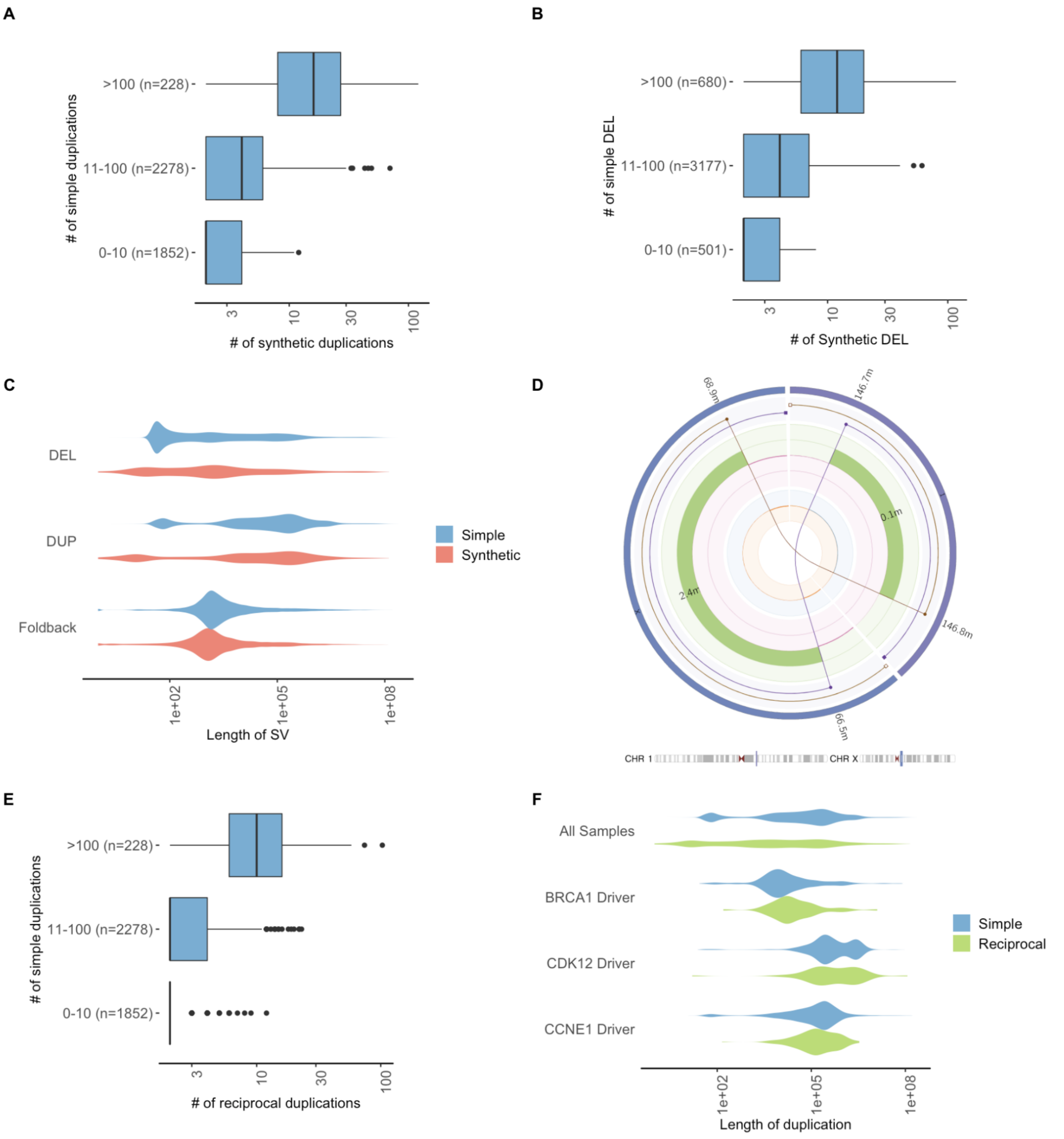


FIGURE S4: Synthetic events and reciprocal duplications, related to Figure 2.

a) & b) Samples with high number of simple duplications (deletions) also have high numbers of 'synthetic' duplications (deletions)

c) Length distribution of deletions, tandem duplications and foldback inversions in Hartwig cohort showing similarity in simple and synthetic length distributions.

d) Example of a predicted reciprocal duplication event involving a pair of translocations from chr 1 to chr X in HMF003502A, a Breast cancer. The brown and purple lines show 2 derivative chromosomes formed by repairing a pair of overlapping breakends on each original chromosome . The rearrangement causes duplication of the regions between the facing breakends on both chromosomes (visible as increased copy number gain in green on the middle track).

e) Samples with high number of simple duplications also have high numbers of reciprocal duplications

f) Length distribution of simple and reciprocal duplications in Hartwig cohort overall and for samples with high confidence BRCA1, CDK12 and CCNE1 drivers showing similarity in length distributions.

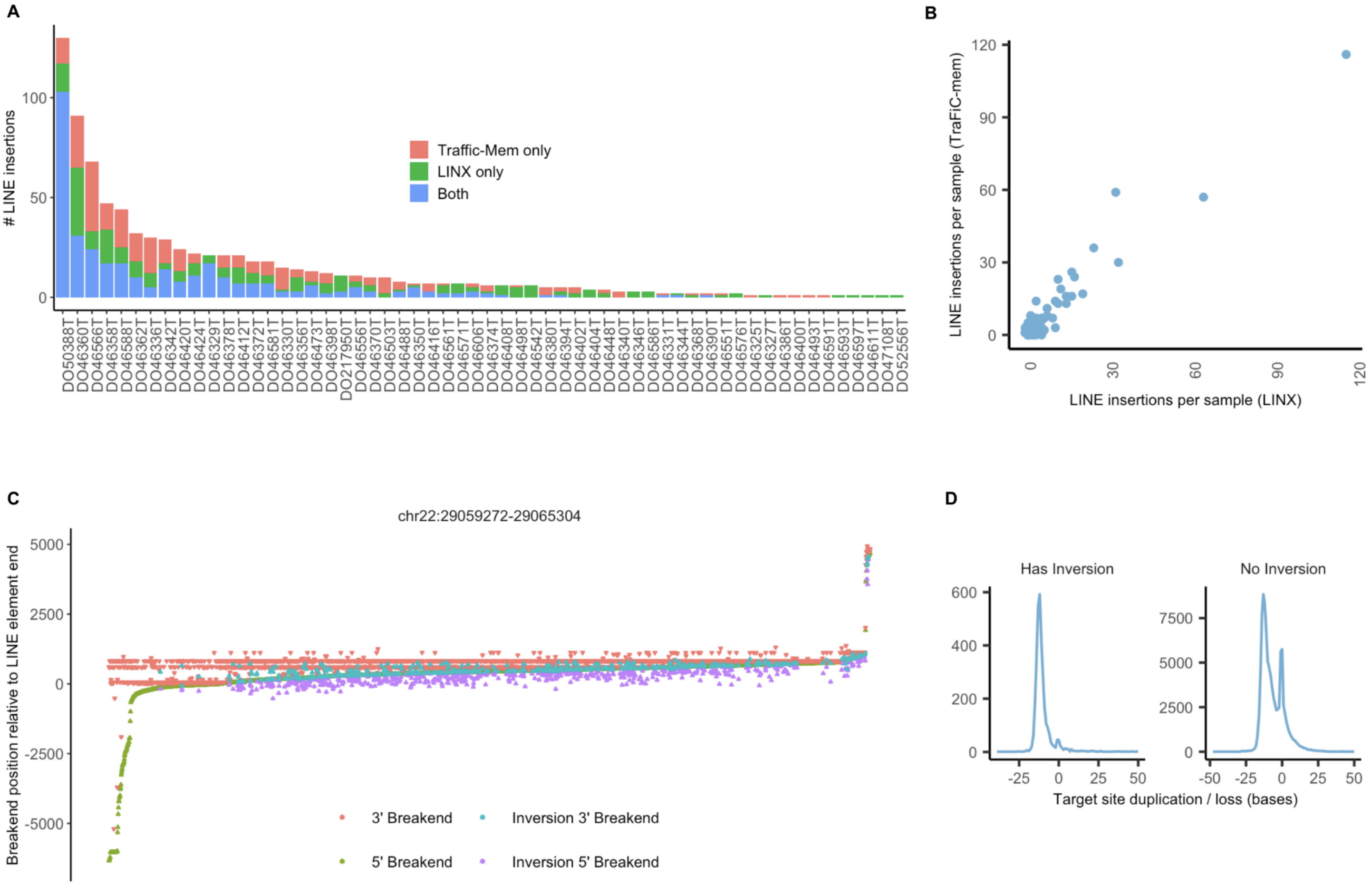


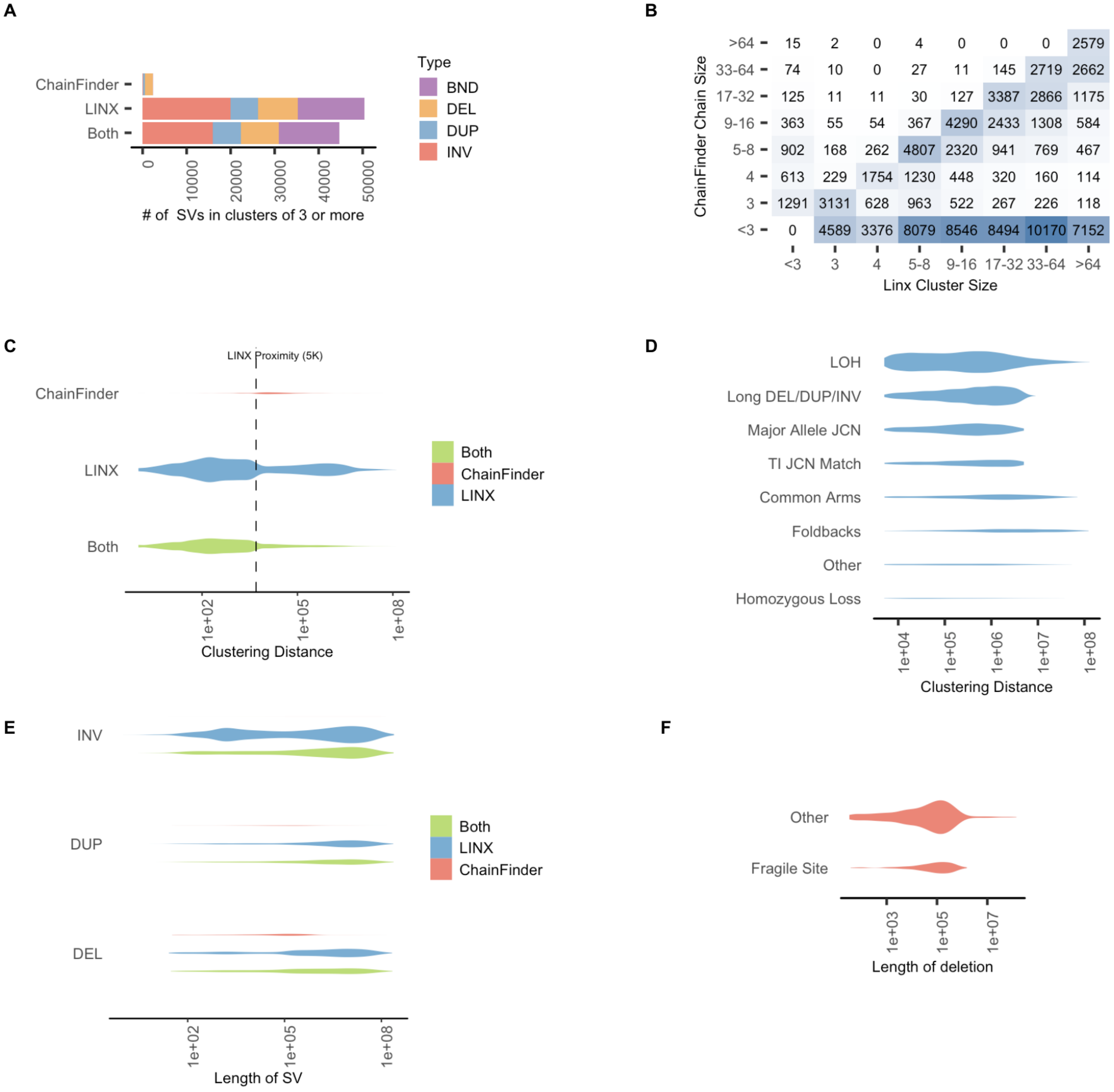
FIGURE S5: LINE insertions validation and analysis, related to Figure 3.

a) Counts of LINE insertions predicted for 75 samples from the PCAWG cohort showing insertions called exclusively in LINX, exclusively in Traffic-Mem and shared in both tools

b) Counts of LINE insertions per sample are highly correlated between LINX and Traffic-Mem

c) Breakends positions for all LINE insertions in Hartwig cohort originating from the frequently somatically activated LINE source element at chr22:29,059,272-29,065,304 relative to the last base of the LINE element in the ref genome. Each column represents one insertion with start and end base of insertion indicated, as well as inversion breakpoints if the insertion contained an inversion.

d) Distribution of target site duplication or loss length for insertions with and without an inversion. Negative values indicate duplication and positive values indicate a loss in bases.



ChainFinder Chain Size	<3	3	4	5-8	9-16	17-32	33-64	>64
>64	15	2	0	4	0	0	0	2579
33-64	74	10	0	27	11	145	2719	2662
17-32	125	11	11	30	127	3387	2866	1175
9-16	363	55	54	367	4290	2433	1308	584
5-8	902	168	262	4807	2320	941	769	467
4	613	229	1754	1230	448	320	160	114
3	1291	3131	628	963	522	267	226	118
<3	0	4589	3376	8079	8546	8494	10170	7152

FIGURE S6: Complex clustering validation, related to Figure 4.

a) Counts of junctions from 1,479 samples clustered into clusters of 3 or more variants by ChainFinder exclusively, LINX exclusively or by both LINX and ChainFinder. More than half of all variants are clustered exclusively by LINX, whereas very few clusters are private to ChainFinder

b) Heatmap of counts of variants clustered into complex clusters by either LINX or ChainFinder showing ChainFinder versus LINX cluster size. Most clusters either have a similar size in both LINX and ChainFinder or are not found at all by ChainFinder

c) Violin plot showing distribution of distance to nearest clustered breakend for all clustered variants by source. The majority of variants clustered in LINX but not chainfinder, have another breakend within 5kb (LINX proximity clustering distance). Nearly all variants with clustering distance >5kb were clustered exclusively by LINX. Area of violin proportional to count of variants.

d) Distribution of distances to nearest clustered breakend for variants clustered exclusively by LINX and for reasons other than proximity. Area of violin proportional to count of variants.

e) Length distribution of same chromosome structural variants clustered by LINX only, Chainfinder only and both. Area of violin proportional to count of variants. Note that ChainFinder private variants are predominantly deletion and duplications <1Mb, whilst LINX private variants follow a similar length distribution to shared variants.
f) Distribution of distances to nearest clustered breakend for deletions clustered by ChainFinder. Area of violin proportional to count of variants. An enrichment of deletions occur in a handful of fragile sites regions which make up <1% of the reference genome.

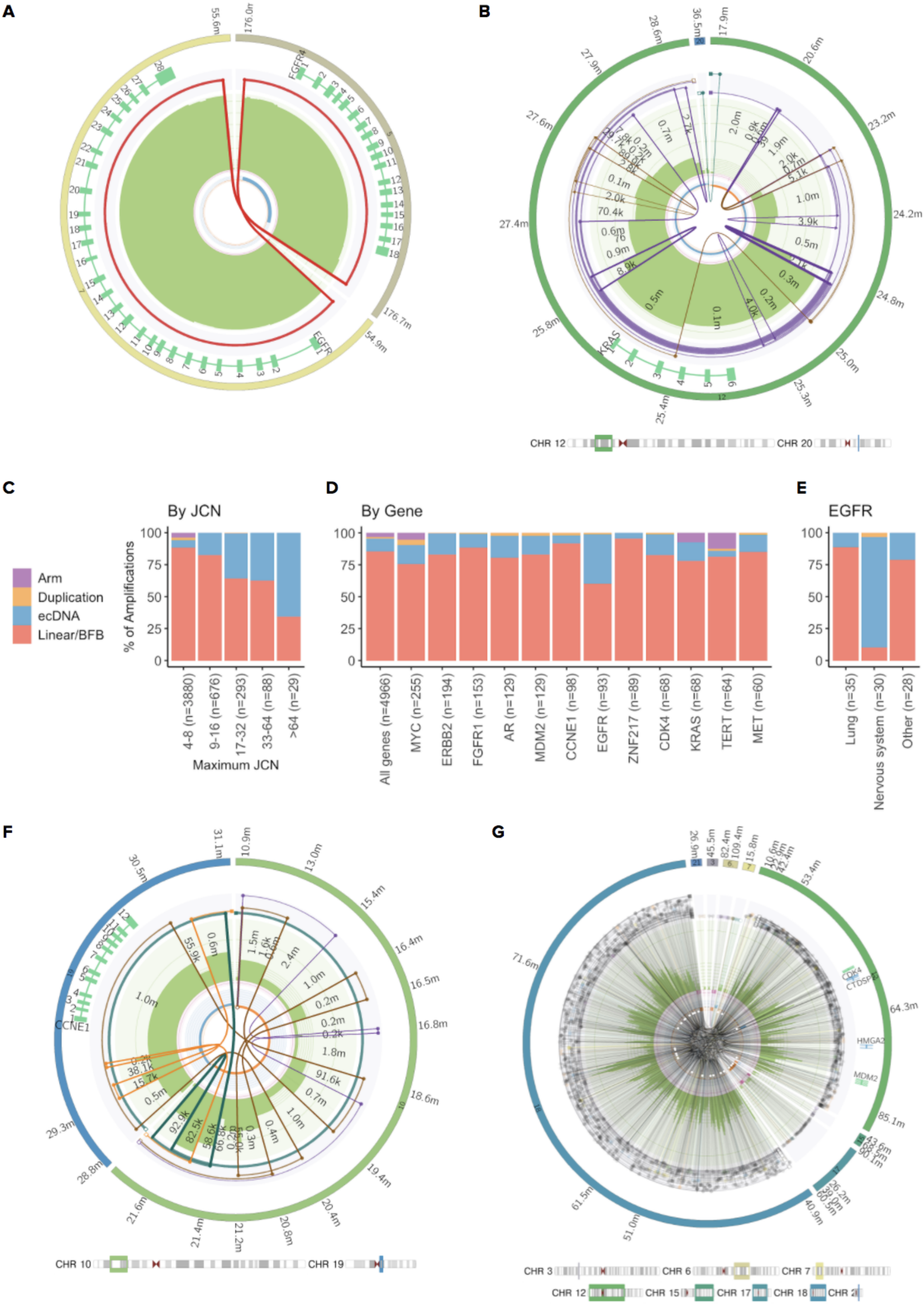

A
55.6m
176.0m
FGFR4
EGFR
176.7m
54.9m
B
28.6m
36.5m
17.9m
20.6m
27.9m
27.6m
23.2m
27.4m
24.2m
24.8m
25.8m
25.0m
25.3m
25.4m
KRAS
CHR 12
CHR 20
C
By JCN
Arm
Duplication
ecDNA
Linear/BFB
% of Amplifications
100
75
50
25
0
4-8 (n=3880)
9-16 (n=676)
17-32 (n=293)
33-64 (n=88)
>64 (n=29)
Maximum JCN
D
By Gene
All genes (n=4966)
MYC (n=255)
ERBB2 (n=194)
FGFR1 (n=153)
AR (n=129)
MDM2 (n=129)
CCNE1 (n=98)
EGFR (n=93)
ZNF217 (n=89)
CDK4 (n=68)
KRAS (n=68)
TERT (n=64)
MET (n=60)
E
EGFR
Lung (n=35)
Nervous system (n=30)
Other (n=28)
F
31.1m
10.9m
13.0m
30.5m
15.4m
16.4m
16.5m
16.8m
CCNE1
18.6m
29.3m
28.8m
19.4m
21.6m
21.4m
21.2m
20.8m
20.4m
CHR 10
CHR 19
G
26.9m
45.5m
82.4m
109.4m
15.8m
53.4m
71.6m
CDK4
CTDSP2
64.3m
HMGA2
MDM2
85.1m
61.5m
51.0m
40.9m
CHR 3
CHR 6
CHR 7
CHR 12
CHR 15
CHR 17
CHR 18
CHR 21

FIGURE S7: High amplification drivers, related to Figure 4.

a) An ecDNA causing amplification of copy number ~60 in HMF002478A, consisting of 2 translocations which amplifies both *FGFR4* on chromosome 5 and *EGFR* on chromosome 7

b) A typical breakage fusion bridge event with 6 foldback inversions leading to high amplification of KRAS in HMF003560A, an ovary tumor. The brown and purple lines show 2 partial and incomplete chains of the foldback event. The green line shows a separate derivative chromosome formed from the same event, clustered due to the breakends bounding the same region of loss of heterozygosity

c)-e) Proportion of amplification drivers which are predominantly caused by each of linear amplifications or breakage fusion bridge,ecDNA, simple duplication or arm level amplifications by maximum JCN (c), driver gene (d) and by tumor type for EGFR amplifications only (e).

f) Complex amplification event affecting CCNE1 in HMF001857A, a non-small cell lung carcinoma. The cluster is not fully resolved by LINX and is instead chained into 4 partial chains (shown in orange, brown, green and purple). The event does not have any foldback inversions, but also lacks regions and breakpoints of uniform high junction copy number expected in ecDNA.

g) A highly complex rearrangement with over 800 junctions in HMF003994A, a malignant peripheral nerve sheath tumor, leading to coamplification of MDM2 and CDK4.

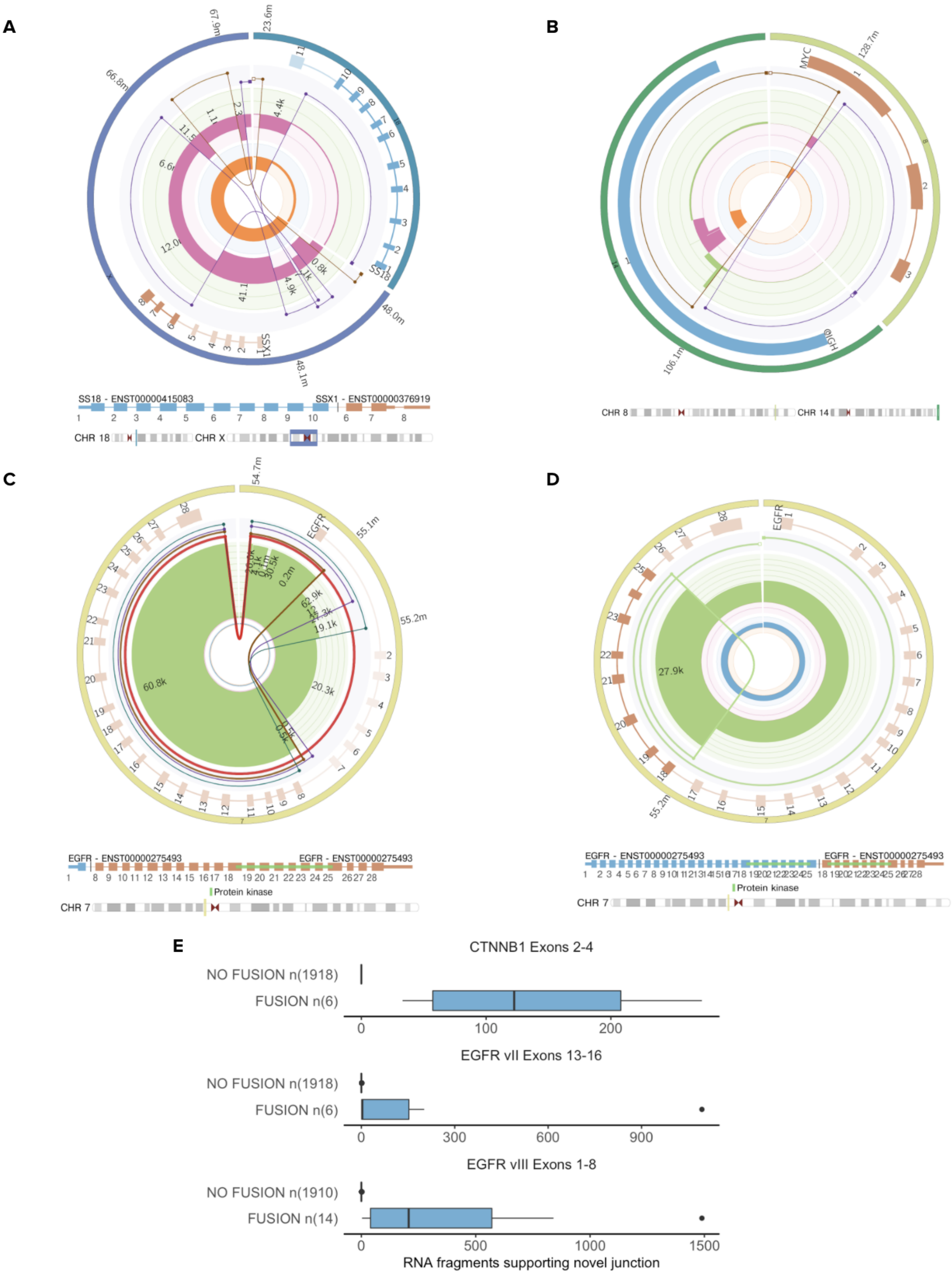


FIGURE S8: Diversity of fusion types, related to Figure 5.

a) A *SS18-SSX1* chained fusion caused by a complex cluster of 6 variants in HMF003579A, a Sarcoma. The predicted derivative chromosome links exons 1-10 in *SS18* to exons 6-8 in *SSX1* via a chain of 2 structural variants.

b) A reciprocal translocation in HMF003019A, a Lymphoid tumor, causing a pathogenic Immunoglobulin rearrangement event by moving the IGH Eμ enhancer adjacent to *MYC*.

c) A double minute in *EGFR* in HMF000649A, a Glioblastoma, formed by a single duplication variant with 3 subsequent trans-phased (and hence independent) subclonal internal deletions of exon 2-7.

d) An *EGFR* kinase domain duplication in HMF004524A, a Lung tumor. Exons 18-25 are duplicated.
e) Counts of RNA fragments with splicing supporting exon deletions for RNA samples with and without the LINX fusion prediction for 3 recurrent exon deletions

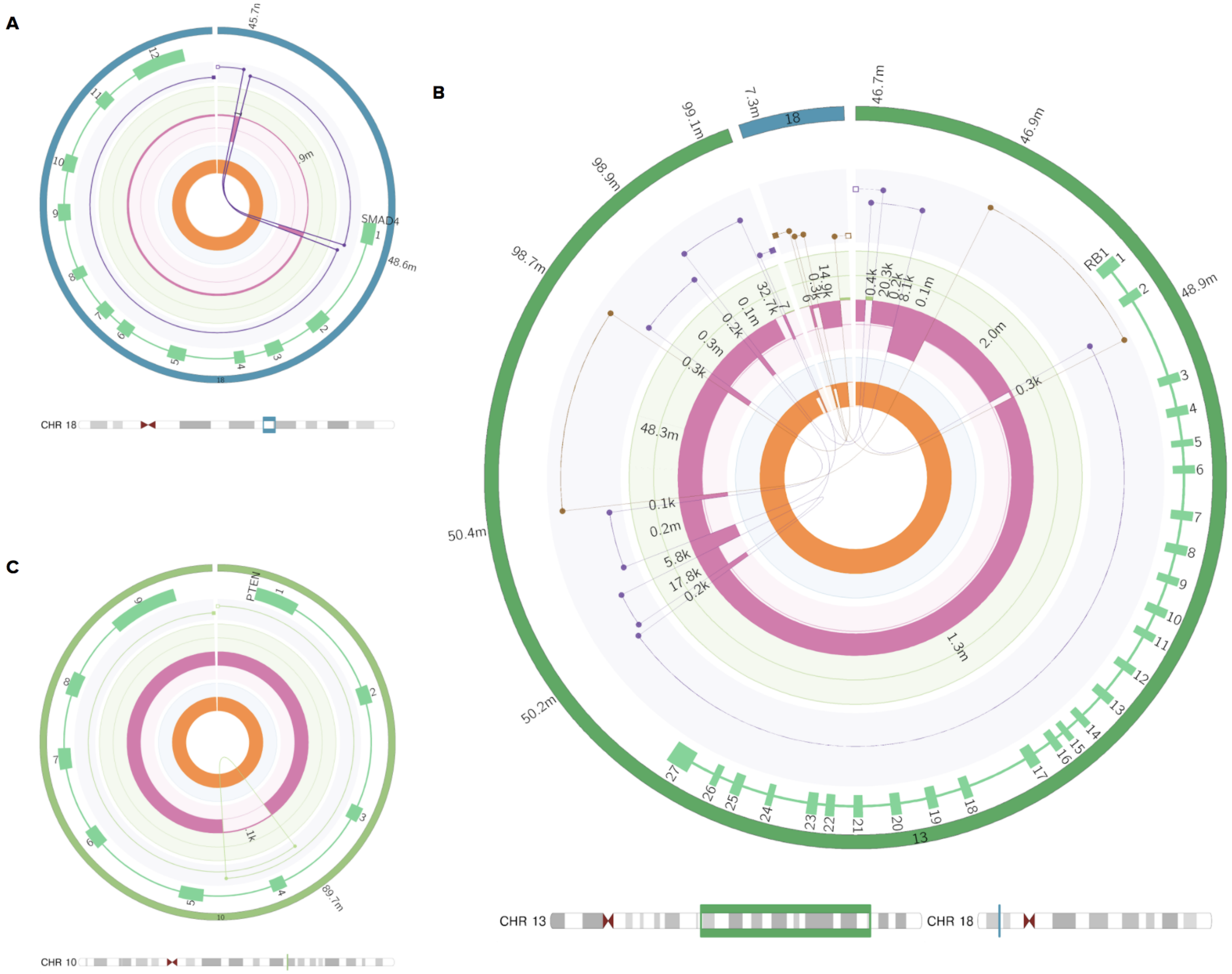


FIGURE S9: Diversity of homozygous disruptions, related to Figure 5.

a) A reciprocal inversion in HMF002869A, a colorectal tumor, which homozygously disrupts *SMAD4* by reversing the direction of exon 1 on the derivative chromosome.

b) A chromoplexy-like event in HMF002065A, a prostate cancer which homozygously disrupts *RB1*. Exons 1 and 2 are predicted to be inserted into chromosome 18 on a separate derivative chromosome (shown in brown) from exons 3-27 which are part of a rearranged chromosome 13 (shown in purple)

c) A homozygous disruption in *PTEN* in HMF001353A, a melanoma, caused by a simple tandem duplication which duplicates exon 4. Since the other parental chromosome is lost, the duplication is predicted to disrupt the last remaining copy of *PTEN*

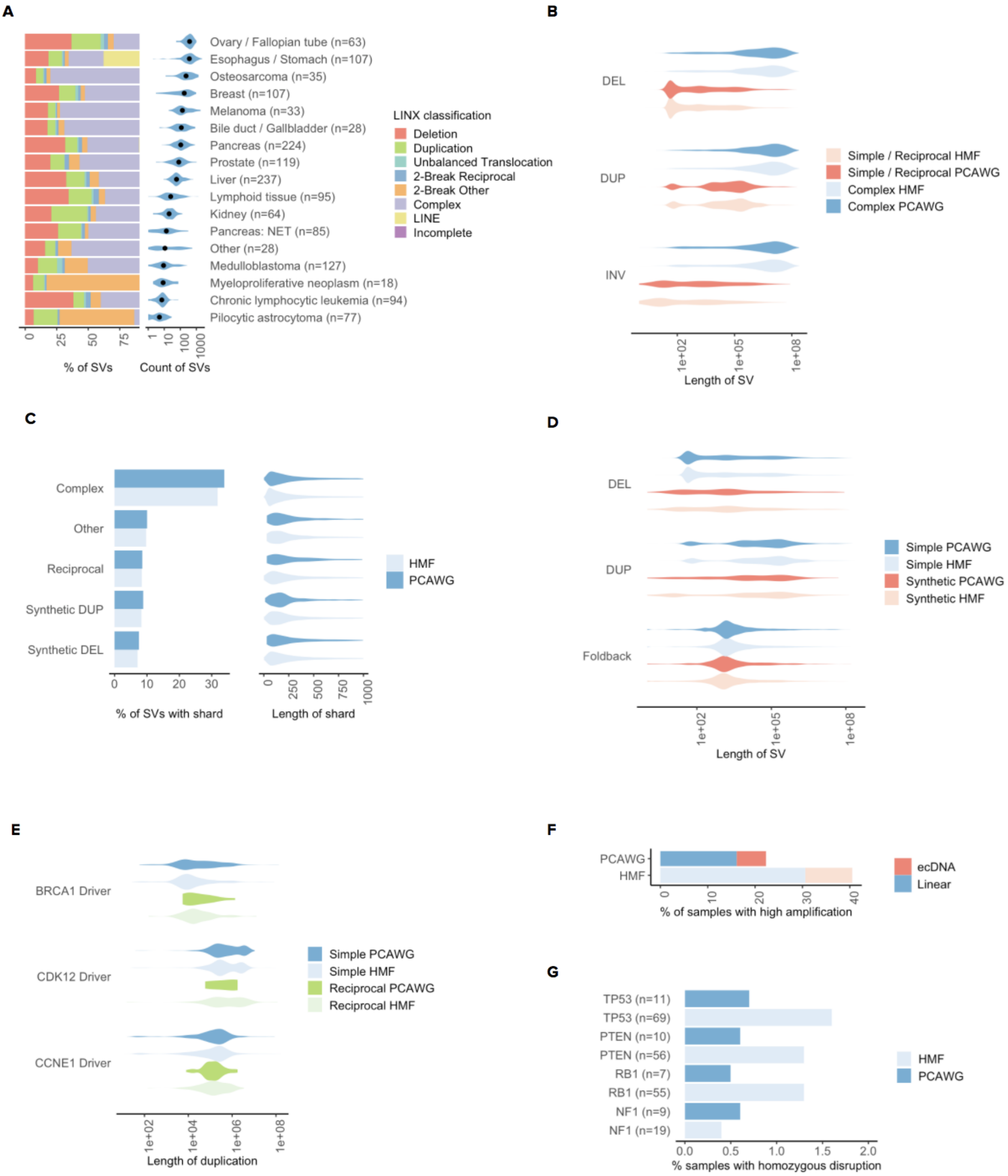


FIGURE S10: Comparison of HMF and PCAWG cohorts, related to Figure 2.

a) LINX classifications of PCAWG samples. Left panel shows relative counts of classifications by tumor type for the PCAWG samples (n=1,541). Right panel shows the number of structural variants per sample grouped by tumour type with the black dots indicating the median values.

b) Length distribution of deletion, duplications and non foldback inversions are highly similar between PCAWG and Hartwig cohorts for both simple rearrangements and complex clusters.

c) Both the proportion of variants with at least one breakend joining a shard of less than 1kb in length (left panel) and the length distributions of shards (right panel) are similar across across the PCAWG and Hartwig cohorts

d) The length distribution of simple and synthetic deletions, tandem duplications and foldback inversions are highly similar across PCAWG and Hartwig cohorts.
e) Length distribution of simple and reciprocal duplications in PCAWG samples with high confidence BRCA1, CDK12 and CCNE1 is similar to the Hartwig cohort.
f) Proportion of samples with high amplification by cohort and mechanism
g) The 4 genes with the highest number of observations of homozygous disruptions are the same for both HMF and PCAWG cohorts

Methods S1: LINX detailed methods description, related to STAR Methods

Overview of key concepts in LINX

LINX terminology and conventions for linking proximate breakends

Assembled, transitive and inferred links

In LINX, 'links' are chromosomal segments flanked by cis phased junctions which are predicted to form part of a single derivative chromosome. Assembled links are those that were called in a single assembly by GRIDSS and are very high confidence somatically phased. Transitive links are not fully assembled by GRIDSS but there is discordant read pair evidence supporting the link as a whole. All other links are inferred, based on proximity, topology and copy number characteristics using the chaining logic described below.

Templated insertions

We have adopted the term 'templated insertion' as has been used previously[1] to describe any piece of DNA which is a templated sequence from a section of the reference genome flanked by breakends on either side inserted elsewhere (either locally or on a remote chromosome) into a chain to form part of a derivative chromosome. The inserted DNA may be either cut (causing disruption) or copied from the template.

'Shards' and 'synthetic' events

A special and very common case of templated insertions we observe are very small templated genomic fragments of up to several hundred bases in length, which are frequently inserted into breakpoints without disruption at the source site for the inserted sequence. These have been observed previously[2] and termed as genomic 'shards'. In LINX we model shards explicitly as short templated insertion lengths of less than 1k bases. These inserted sequences can make simple events such as deletions and tandem deletions appear to have complex topologies. For example, if we have a simple short deletion with a shard inserted, and the templated sequence of the shard is from another chromosome the deletion now presents notionally as a chained pair of translocations. Where more than 1 shard is inserted, the complexity can grow even further. LINX simplifies events that could be explained as a 1 or 2-break cluster with shards and marks the cluster as 'synthetic'. Figure S1 shows a number of examples of synthetic events with the shards marked.

Deletion bridges, anchor distance & overlapping deletion bridges

We use the term 'deletion bridge' as defined previously[3] to refer to sections of DNA loss between 2 breakpoints on the same paternal chromosome that are fused to other segments of the genome. GRIDSS provides an anchor support distance for each structural variant breakend which is the number of bases mapped to the reference genome at that breakend as part of the assembly contig, which is typically in a range from 29 bases (the minimum anchor distance for GRIDSS to be able to call) up to approximately 800 bases for short read sequencing. Any other breakend that falls within this anchor distance cannot be 'cis' phased with the variant as the contig was able to be mapped past the breakend and the 2 breakends are deemed to be 'trans' phased. Trans breakends within this distance range are common in cancer. One possibility is that the breakends could occur on the other paternal chromosome, but this highly unlikely as there is no reason to expect 2 different paternal chromosomes to both be damaged within a few hundred base region. Much more likely is that when the double stranded break occurred, that there was significant overlap between the break locations on the 2 strands and the shorter strand of each overlapping break end has been repaired prior to fusing with other regions of the genome . This is highly analogous to a deletion bridge except with small sections of replication of DNA instead of loss. LINX uses the term 'overlapping deletion bridge' to describe this breakend topology.

Copy number conventions

LINX determines the number of absolute copies of each rearrangement junction in a sample, and terms this as the “junction copy number” (JCN). PURPLE SV output provides both a raw estimate of the JCN (estimated from the purity adjusted VAF of the junction) as well as the change in copy number observed at each breakend. LINX uses both the raw estimate and the copy number change to predict both a JCN point estimate and uncertainty for each rearrangement junction.

Overview of event classification system in LINX

LINX attempts to classify all variants into a set of consistent events, i.e. events that transform the genome from one stable configuration into another. LINX classifies all events with one or two junctions and groups events with 3 or more junctions as ‘COMPLEX’.

A key assumption in LINX is that each derivative chromosome arm in a stable configuration must connect a telomere to a centromere (since centromere to centromere joins will cause unstable breakage fusion bridge and telomere to telomere joins will have no centromere and will be ultimately lost during stochastic mitosis processes). A special case is allowed in highly restricted circumstances for double minute chromosomes which are circular and have no telomere or centromere but are highly positively selected for. This assumption means that variants such as a lone head to head or tail to tail inversion are considered incomplete, and in these cases we intensively search for other variants which may have occurred concurrently and could restore a stable configuration. Because of limitations of both input data accuracy and completeness and our clustering and chaining algorithm, many COMPLEX clusters will not be fully resolved to a stable configuration although it is assumed that such a resolution exists. Furthermore, we have a number of residual 1 and 2 clusters (eg. a lone inversion) which are inconsistent cannot be accurately clustered and hence we classify them as INCOMPLETE.

Ultimately we classify each cluster into 1 of 7 major event categories:

Event Category	Description
SIMPLE	Single junction cluster which forms a local deletion, tandem duplication or unbalanced translocation
RECIPROCAL	Reciprocal inversion or translocation events forming from 2 concurrent breaks interacting with each other
TEMPLATED INSERTION	DEL or DUP or unbalanced translocation (‘chain’) with templated insertion
INSERTION	SV that are formed by the insertion of a templated piece of DNA normally via either a mobile element insertion or virus.
DOUBLE_MINUTE	Any 1 or 2 variant cluster where all variants form part of an ecDNA ring
COMPLEX	Clusters with 3 or more variants that cannot be resolved into one of the above categories
INCOMPLETE	1 or 2 breakpoint clusters which are inconsistent, but cannot be clustered further OR clusters which are inferred from copy number changes only

A brief overview of each of the non SIMPLE categories is given below:

Reciprocal events

Linx models reciprocal 2-break junction clusters as events that could be caused by the interaction of 2 simple local concurrent breaks which would normally form deletes and tandem duplications. Depending on whether the breaks are on the same or different chromosomes this forms reciprocal inversions or reciprocal translocations respectively. Note that in the translocation case, if one side of the reciprocal event is subsequently lost either before or after repair, then we will instead observe an unbalanced translocation .

The possible geometries for reciprocal events supported by LINX are explained in the table below and drawn in figure S2:

Interacting Break Types	Same chromosome (Inversion)	Translocation
Concurrent double stranded breaks	**RECIP_INV** - 2 facing inversions with outer breakends overlapping	**RECIP_TRANS** - 2 translocations forming deletion bridges on both arms.
Concurrent tandem duplications	**RECIP_INV_DUPS** - 2 facing inversion with inner breakends overlapping	**RECIP_TRANS_DUPS** - 2 translocations with facing breakends on both arms
Tandem Duplication + Double Stranded Break	**RECIP_INV_DEL_DUP -** inversion enclosing inversion with opposite orientation	**RECIP_TRANS_DEL_DUP -** 2 translocations forming a deletion bridge on one arm and facing breakends on other arm

A facing pair of foldback inversions (FB_INV_PAIR) is also classified as a reciprocal, although the mechanism for forming this structure is unclear. It is possible that many of these events are formed from a breakage fusion bridge event but have not been properly clustered with a resolving break junction which may be distant in a breakage fusion bridge scenario.

Templated insertions

For the 4 reciprocal event cases above involving duplication (ie. RECIP_INV_DUPS, RECIP_INV_DEL_DUP, RECIP_TRANS_DUPS & RECIP_TRANS_DEL_DUP), the same junctions can be alternately chained to form a single derivative chromosome with a templated insertion (see figure S2). LINX gives precedence to the reciprocal interpretation, but if any of the duplicated segments bound a telomeric or centromeric loss of heterozygosity, the reciprocal interpretation is implausible

A deletion and duplication can together also form either a duplication or deletion with templated insertion structure (figure S2) identical to the 2 inversion case but with the inserted segment in the opposite orientation. Unlike inversions, simple deletions and tandem duplications are consistent standalone events and are common genomic events so some of these structures may be clustered incorrectly where separate DEL and DUP events are highly proximate or overlapping by chance.

Insertions

An insertion event is modelled by LINX as a pair of structural variants which inserts a section of templated sequence from either another part of the genome WITHOUT disruption to the DNA at the source location OR from an external sequence such as an insertion from a viral genome.

The most common class of insertion in tumor genomes by far are mobile element insertions, which are not typically active in the germline, but can be highly deregulated in many different types of cancer. Mobile elements insertions frequently insert short sequences of their own DNA sequence and templated segments from adjacent to the source LINE region, with sometimes many segments from the same source location being inserted at multiple locations around the genome[4]. Activated LINE can also cause SINE and pseudogene insertions. LINE insertion source breakends can be often difficult to map correctly on both ends, since they typically involve a repetitive LINE motif at the start of the insertion element and a poly-A section at the end of the inserted section. LINX uses a combination of previously known active LINE source region information and identification of both the local breakpoint structure and POLY-A sequences to classify both fully and partially mapped breakpoints as LINE insertions.

Double minute

Any 1 or 2 variant cluster which is predicted to form a closed loop by LINX without a centromere is resolved as a ‘double minute’. All variants must form part of the ecDNA to be classified as event type double minute, although ecDNA may also occur as part of a complex cluster. An exception is made for a simple DUP double minute clustered with an enclosing DEL, which is classified as double minute despite the DEL not being a part of the ecDNA structure. Complex clusters may also contain double minutes.

Complex events

COMPLEX events are defined in LINX as clusters with 3 or more variants that cannot be resolved into either a simple or synthetic type of insertion, DEL, DUP or 2-break event.

COMPLEX events may be formed by any combination of non-mutually exclusive processes including multiple concurrent breaks, replication prior to repair, breakage fusion bridge processes. Local topology annotations in LINX are intended to shed light on these complex processes.

Incomplete events

There are a number of possible configurations which are not ‘COMPLEX’ by the above definition since they are formed from 1 or 2 SVs, but lead to inconsistent genomes or involve single breakends. For these clusters there is assumed to be missing SVs, potential false positive artifacts or under clustering and they are marked as INCOMPLETE.

INCOMPLETE includes but is not limited to the following configurations:

- Lone inversion
- Lone single breakend
- Lone inferred breakend
- Any 2-break junction cluster with a single or inferred breakend that cannot be resolved as LINE or inferred as a synthetic.
- Any 2-break junction cluster which cannot be chained OR resolved as either a LINE, synthetic,templa or reciprocal event

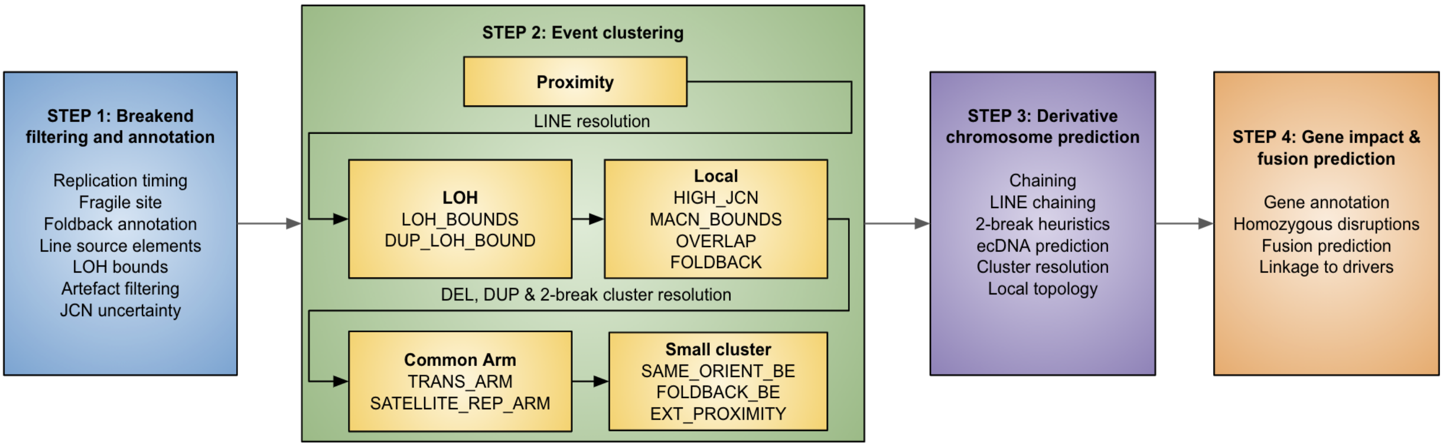
Clusters of 2 inferred breakends are also classified in this category. Many of these are likely artifacts due to residual large scale GC biases affecting coverage unevenness in the sequencing data and false positive CNA calls.

LINX algorithm

There are 4 key steps in the LINX algorithm:

- Annotation of genomic properties and features
- Clustering of SVs into events
- Chaining of derivative chromosomes
- Gene impact and fusion prediction

The following schematic outlines the overall workflow in the LINX algorithm. Each step is described in detail below



1. Annotation of genomic properties and features

To help resolve and characterise events, LINX first annotates a number of genomic properties:

Externally sourced genomic annotations

Each breakend is first annotated with the following information from external sources

- Whether it is in a known fragile site[5]
- Whether it is in a known LINE source region[4]
- The replication timing of the breakend[6]

Identification of foldback inversions

Foldback inversions are important structural features in the genome since they are a hallmark of the breakage fusion bridge process. LINX use foldbacks in a number of ways in both the clustering and chaining algorithms, and they can be objectively identified independently of the clustering and chaining so it is useful to identify them upfront. We perform a genome wide search for both simple foldback inversions and chained foldback inversions which are disrupted by the insertion of a shard. A pair of breakends is marked as forming a foldback if they meet the following criteria:

- the breakend orientations are the same and are consecutive (ignoring any fully assembled interim breakends) and after allowing for overlapping deletion bridges on both ends, specifically both:
    - The outer breakend may be overlapped by a variant within it's anchor distance
    - The inner breakend may not have a facing breakend within it's anchor distance
- the breakends belong to a single inversion or are linked by an assembled or short chain (<= 5K bases)
- A single breakend where the other end of the structural variant is assembled to itself via a chain
- Neither breakend forming the foldback is linked via assembly to another breakend.

Identification of suspected LINE source regions

LINE source regions are also important genomic features and are modelled in LINX as regions of ~5000 bases which LINX suspects are the source for templated LINE insertions. LINE driven mobile element insertions are common in cancer genomes and typically present as a pair of balanced SVs representing a templated sequence from around the source element with a poly-A tail inserted into random locations in the genome although favoring a A|TTTTT motif[4] for the insertion site with no net copy number change at either source or insertion site. However, due to both the repetitive nature of the LINE source regions and the difficulty of accurately sequencing across the poly-A tail, one or both of the SVs that make up the insertion may be mapped as a single breakend (failing to uniquely map on the other side) OR be missed altogether. Since the lone remaining breakend can be mistaken as an unbalanced translocation, it is important to correctly identify it as a LINE insertion. This picture can be complicated even further by the fact that many overlapping fragments from a single source location may be copied to many different locations in the same genome, each potentially with one or both sides incompletely mapped.

Although we already annotate 124 'known' mobile LINE source regions which have been previously discovered[4], there are many more potential mobile LINE source regions which may be less common in the population or more rarely activated. We look exhaustively for likely LINE source regions in each individual tumor genome, by looking for both the characteristic poly-A tail of mobile element insertions and the local break topology structure at the insertion site. We define a poly-A tail as either at least 16 of the last 20 bases of the sequence are A or there is a repeat of 10 or more consecutive A or within the last 20 bases of the insert sequence. The orientation of the breakend relative to the insertion can help distinguish between the source and insertion site for a mobile element. At the mobile element source site, the poly-A tail positive oriented breakends will have the poly-A at the start of the insert sequence, or poly-T at the end of the insert sequence for negative oriented breakends (if sourced from the reverse strand). Conversely at the insertion site, negative oriented breakends will have poly-A tails at the end of the insert sequence and positive oriented breakends have poly-T at the start of the insert sequence (if inserted on the reverse strand)

Each breakend will be classified as being a 'Suspected' LINE source region if any of the following conditions are met:

- there are 2+ breakends within 5kb with poly-A/poly-T tails with expected orientations for a source site.
- there are 2+ translocations or local junctions > 1M bases which are not connected at their remote end to a known LINE site within 5kb with at least one not forming a deletion bridge of < 30 bases AND at least one breakend within 5kb having a poly-A tail with expected orientation for a source site
- we find at least 1 translocation or local junction >1M bases with it's remote breakend proximity clustered with ONLY 1 single breakend and forming a deletion bridge of < 30 bases AND EITHER the junction has a poly-A / poly-T tail with the expected orientation of the source site OR the remote single breakend has a poly-A/poly-T tail with expected orientation for an insertion site.

The suspected LINE source region is also checked that it is not a potential pseudogene insertion by checking that there is no deletion within 5kb of the suspected source element that matches an exon boundary at both ends. Both known and suspected source elements have special logic applied in the clustering phase.

Identification of LOH boundaries

LINX also identifies each pair of breakends flanking regions of Loss of Heterozygosity (LOH), restricted to cases where there is no subset of the region with homozygous loss that is not caused by anything other than a simple deletion. This is a useful annotation as since an entire paternal allele is lost for the whole distance between these 2 breakends (and since there is no homozygous loss we know it is the same allele lost at both ends, not two overlapping losses) then the structural variants are very likely to have occurred at the same time.

Note that an uninterrupted deletion or tandem duplication cannot theoretically form an LOH boundary with another variant and hence these are excluded from LOH boundaries.

Long DEL and DUP length calculation

Shorter deletions and tandem duplications are found frequently as standalone events in the tumor genome, but longer standalone deletions and duplications are relatively rare and when they do occur are often associated with more complex events. The following method is used to determine a characteristic length threshold for each sample for what is considered a 'long' DEL or DUP.

- find all DUPs and DELs on arms with no inversions (inversions are used as a proxy for the presence of complex events)
- Set LONG_DUP_LENGTH to the length of the longest DUP excluding the 5 longest DUP lengths (normalised for the proportion of arms without inversions). Min =100k, Max = 5M
- Set LONG_DEL_LENGTH to the length of the longest DEL excluding the 5 longest DEL lengths (normalised for the proportion of arms without inversions). Min =100k, Max = 5M

The threshold is subsequently used in clustering rules by LINX.

Estimation of JCN and JCN uncertainty per SV

We know that by definition JCN must equal the copy number change at start breakend and the copy number change at end breakend for each and every structural variant. However, frequently the JCN and copy number change at start and end may not match up for one of several reasons including measurement error, false positive SV artifacts and false negative SV calls adjacent to one or both of the breakends for the structural variant.

To allow for accurate chaining we would like to have a single consolidated JCN estimation and an idea about the uncertainty in the JCN for each SV. Since the distribution of the errors in our measurements are unknown and fat tailed due to potentially missing and false positive data, we create a simple model for a reasonable estimation of the likely JCN range.

We use 3 steps in the process:

1. Estimate an uncertainty for copy number change at each breakend

For this use the principle that the uncertainty in copy number change is driver primarily by the uncertainty in the copy number of the least confident adjacent copy number region which in turn is driven primarily by the number or read depth windows used to estimate the length of the adjacent regions.

Hence we use the following formula to calculate a copy number uncertainty

> CNChangeUncertainty = MAX(maxAdjCN * BaseRelUncertainy[0.1],BaseAbsUncertainty [0.15])+
> MAX(AdditionalAbsUncertainty [0.4],AdditionalRelUncertainty[0.15]*maxAdjCN)
> /SQRT(max(minAdjDepthWindowCount,0.1))

If the minAdjacentDepthWindowCount = 0, then this means the segment is inferred by the raw JCN in PURPLE already and no copy number estimate is calculated.

For the special case of foldback inversions, if the flanking depth window counts are both higher than the internal depth window count, a single copy number change observation of half the combined copy number change is made with the confidences determined from the flanking windows

2. Estimate an uncertainty for raw JCN

The raw JCN of the SV is already estimated in PURPLE by multiplying the purity adjusted VAF of the SV by the copyNumber at each breakend. The VAF estimate depends ultimately on the measured readcount of supporting tumor fragments which is a binomial distribution.

To estimate the uncertainty in the VAF, we estimate the 0.5% and 99.5% confidence intervals of the true read count from the observed read count and then calculate the JCN uncertainty as half the relative range of the confidence interval. We also add half the minimum of the 2 breakend copy number uncertainties to reflect the copyNumber impact on uncertainty. This gives the formula:

> JCN Uncertainty = JCN * (ReadCountUpperCI-ReadCountLowerCI) / 2 / ObseverdReadCount + 0.5 *
> min(CNChangeUncertaintyStart,CNChangeUncertaintyEnd)

3. Average the 3 JCN predictions and estimate a consolidated uncertainty

Weight the observations by the inverse square of their estimated uncertainties:

> consolidatedJCN = SUM[Observation(i)*(1/Uncertainty(i)^2)] / Sum[1/Uncertainty(i)^2]

The combined uncertainty is estimated as the square root of the weighted sum of squares of the difference between the final JCN estimate and each individual estimate, but capped at a minimum of half the input uncertainty. Ie.

```
consolidatedUncertainty = SQRT( countObservations / (countObervations-1) *
        SUM[1/Uncertainty(i)^2*(MAX(Observation(i)-consolidatedJCN,Uncertainty(i)/2))^2] /
        Sum[1/Uncertainty(i)^2] )
```

Artifact filtering

Prior to clustering and chaining, LINX applies additional artifact filtering, since despite applying stringent filters in both PURPLE and GRIDSS upstream, we may still find a number of false positive artifacts in our data. False positive artifacts are typically either SVs with little or no copy number support (ie copy number change at both breakends < 0.5) or inferred SV breakends from the copy number analysis. Unfortunately these can be difficult to distinguish from bonafide subclonal variants with ploidies of 0.5, and from genuine clonal variants where we have missed an offsetting SV call which netted out the copy number change.

To remove residual artifacts, but preserve genuine subclonal variants, we limit filtering to 3 very specific situations which strongly appear to be artifactual in our data:

- **Short foldback Inversions (<100 bases) unsupported by copy number change -** Typically foldback inversions range from several hundred to several thousand bases. However, we also frequently find many very short foldback inversions in highly damaged samples. Across the cohort as a whole we find these to overwhelmingly have low junction copy number and with little copy number support. Hence we mark foldback inversions <100 bases in length as artifacts if both start and end copy number change < 0.5 or VAF < 0.05 at both ends.
- **Isolated translocations and single breakends unsupported by copy number change at both breakends -** These also common artifacts and similar to foldback inversions, we find an elevated rate of low junction copy number variant calls unsupported by copy number change indicating likely artifacts. We filter if both breakends of a translocation are > 5000 bases from another variant AND both breakends have copy number change <0.5 AND the insert sequence neither has a poly-A tail nor insertSequenceRepeatClass = 'LINE/L1' (which may indicate a Line insertion).
- **Neighbouring inferred breakends with opposite orientation and matching or overlapping JCN uncertainty** - Residual GC bias and other forms of read depth noise can cause many inferred segments to be called. These are unhelpful in chaining and are resolved in pairs of offsetting variants as artifacts

All of the filtered variants will be marked as resolved type = ARTIFACT and will be restricted from any subsequent clustering.

2. Clustering of structural variants into events

LINX uses a clustering routine to classify events. All SVs within a sample are grouped into clusters in 7 steps:

- Proximity clustering
- Resolution of LINE clusters
- LOH clustering
- Local clustering
- Resolution of simple events
- Common arm clustering
- Incomplete and small complex cluster merging

Proximity clustering (PROXIMITY)

Any SV with a breakend within the specified proximity_distance (default = 5Kb) of another SV's breakend causes the SVs to be clustered. An exception is made for overlapping DELs, which are split out into separate clusters, on the assumption that overlapping DELS must by definition be trans-phased.

Resolution of LINE clusters

LINX performs early resolution of mobile element insertions as insertions typically cause translocations that may inadvertently be clustered with other variants, particularly in tumors with highly deregulated LINE activation.

LINX resolves a cluster as type LINE if any of the following conditions are met:

- It contains a suspected LINE source region AND (the cluster has <=10 variants OR at least 50% of the non single or inferred junctions in the cluster have a known or suspected breakend)
- Every non single or inferred breakend variant in the cluster touches a KNOWN LINE source region AND at LEAST one of the variants is a translocation

Due to the low mappability of the LINE source regions, frequently we will also identify LINE insertions only as single breakends as the insertion site. We therefore also resolve a cluster as LINE insertion if it contains either :

- 2 single breakends only, forming a deletion bridge (less than +/-30 bases) with [one side having a poly-A/poly-T tail with the expected orientation of an insertion site OR one of them has insertSequenceRepeatClass = 'LINE/L1']
- 1 single breakend and 1 inferred breakend only and the single breakend has insertSequenceRepeatClass = 'LINE/L1' or a poly-A/poly-T tail with the expected orientation of an insertion site
- 1 single breakend with insertSequenceRepeatClass = 'LINE/L1' or a poly-A/poly-T tail with the expected orientation of an insertion site and copy number change < 0.5 / 15%

LINE clusters are excluded from all subsequent clustering rules.

LOH Clustering

Loss of heterozygosity bounds (LOH_BOUNDS)

The 2 breakends forming the bounds of any LOH region not disrupted by a homozygous deletion are clustered together, reflecting the fact that both ends of the lost paternal chromosome must have been lost at the same time.

For LOH regions which are disrupted by 1 or more homozygous deletions, both paternal chromosomes are presumed to have been deleted in separate events and one deletion may enclose the other deletion OR they may overlap each other. In the enclosing case, for homozygous deletions which are in a region where the surrounding LOH bounds which form a simple DEL or are already linked or extend to at least the full arm, then we cluster the 2 homozygous deletion bounds. Conversely if all the homozygous deletions inside a LOH region are simple DEL or are already linked then we can cluster the 2 LOH region bounds. Finally, in the overlapping case, if all the overlapping deletion bounds are from simple DELs or are already linked except for 2 breakends, then link the remaining 2 breakends.

Note that variants with breakends that bound an LOH are not permitted to link with breakends in the bounded LOH region via any subsequent rule, since by definition they are expected to be on the other paternal chromosome.

Chaining bounds for DUP variants causing LOH (DUP_LOH_BOUNDS)

No breakend in a cluster can chain across an LOH which has been caused by a breakend in the same cluster. Hence if the other breakend of a DUP type variant bounding an LOH can only chain to only one available (not assembled, not LINE) breakend prior to the LOH, then we cluster the DUP and the other breakend.

Local Clustering

High JCN (HIGH_JCN)

Merge any pair of breakends which both have JCN> max(5, 2.3x the adjacent major allele copy number) that face each other that are also joined by a continuous region of majorAlleleCN > 5

Major allele copy number bounds (MACN_BOUNDS)

The major allele copy number of a segment is the maximum copy number any derivative chromosome which includes that segment can have. Hence a breakend cannot chain completely across a region with major allele copy number < JCN of the breakend, or partially across the region with a chain of JCN more than the major allele.

Therefore any breakend is clustered with the next 1 or more facing breakends (excluding LINE & assembled & simple non overlapping DEL/DUP) IF the major allele copy number in the segment immediately after the facing breakend is lower than the breakend JCN, after discounting facing breakends in the original cluster. In the case of there being more than 1 facing breakend, the highest JCN breakend is clustered and the process repeated. This clustering is limited to a proximity of 5 million bases and bounded by the centromere, since although more distal events on the same chromosome may be definitely on the same derivative chromosome, this does necessarily imply they occurred concurrently.

Local Overlap (OVERLAP)

Merge any clusters or SVs where each cluster has either an inversion or a DEL exceeding the LONG_DEL_LENGTH, or a DUP exceeding the LONG_DUP_LENGTH and they overlap or enclose each other AND the 2 variants have at least one pair of breakends not more than 5M bases apart that either face each other or form a deletion bridge.

Foldbacks on same arm (FOLDBACK)

Merge any 2 clusters with a foldback on the same chromosomal arm.

Resolve DEL, DUP and consistent 2-break junction events

At this step, DEL, DUP and reciprocal clusters are resolved to prevent them from over clustering with later clustering rules. Specifically, the following types of events are resolved:

- **Deletions** - Simple deletions less than the LONG_DEL_LENGTH
- **Tandem duplications** - Simple duplications less than the LONG DUP_LENGTH
- **Reciprocal events and 2-break junction templated insertions -** Pairs of overlapping inversions or translocations where the deletion bridge and/or overlap at both breakends are less than the LONG_DEL_LENGTH / LONG_DUP_LENGTH and neither deletion bridge or overlapping region contains a non simple DEL/DUP breakend.

We also allow for synthetic variants of these events to be resolved, where the variants create a DEL, DUP or reciprocal event with the same geometry but with one or more shards inserted. All of these simple and synthetic clusters types are excluded from all subsequent clustering rules.

Common arm clustering rules

Translocations with common arms (TRANS_ARM)

Merge any 2 unresolved clusters if they touch the same 2 chromosomal arms. SVs which link the 2 arms but are in a short templated insertion (< 1kb) are ignored.

Single breakends on same arm with matching satellite repeat type clustering (SATELLITE_SGL_ARM)

Where complex events touch satellite repeats we frequently find many single breakends on the same chromosome with links to the same type of repeat. In particular this can occur when shattering events include complex focal scarring in centromeric regions leading to many unresolved single breakends

We therefore merge any cluster with less than or equal to 1 non single breakend and non inferredbreakend with any other cluster which contains a single breakend on the same chromosome arm with matching repeat class or type for the following cases:

- RepeatClass = 'Satellite/centr’ (centromeric)
- RepeatType = '(CATTC)n’ (satellite repeat type)
- RepeatType = ‘(GAATG)n’ (satellite repeat type)
- RepatType = ‘HSATII’ (pericentromeric)

To protect against false positives and joining complex clusters that both touch repeats, but otherwise don’t appear to overlap, we avoid clustering 2 clusters which already have multiple non single breakends.

We don’t cluster other common sequences such as telomeric sequences, Sine/Alu or LINE/L1 as these tend to be associated with genome wide insertion patterns rather than specific clusters which touch a repetitive region.

Incomplete and small complex cluster merging

These rules are implemented to merge small unresolve with 3 or less variants to other unresolved clusters with an arbitrary cluster size where the location and orientation of proximate or overlapping breakends between the 2 clusters indicates that they may be linked.

Breakends straddled by consecutive same orientation breakends (SAME_ORIENT_BE)

Merge any non-resolved breakend to a cluster which straddles it immediately on both sides with 2 breakends facing the same direction, and where the facing breakends have matching JCN.

Breakends straddled by foldbacks (FOLDBACK_BE)

Merge any non resolved breakend into a cluster which has 2 different foldbacks straddling it immediately on both sides and at least one of the foldbacks faces the breakend.

Extended chainable proximity for complex and incomplete events (EXT_PROXIMITY)

Merge any neighbouring non resolved clusters that are within 5M bases and which have facing flanking breakends on each cluster which could form a templated insertion with matching JCN. In the case of a foldback the JCN of the facing breakend is also permitted to match 2x the JCN.

3. Chaining of derivative chromosomes

A chaining algorithm is used to predict the local structure of the derivative chromosome within each cluster. The chaining algorithm examines each cluster independently and considers all possible paths that could be made to connect facing breakends into a set of continuous derivative chromosomes.

Overview of chaining model

Chains in LINX are a walkable set of linked breakends with a common JCN. Initially each cluster begins with 1 chain for every SV in the cluster. LINX iteratively makes 'links' between chain ends, resolving 2 of the chains into 1 combined chain and determining a combined JCN and JCN uncertainty for the new combined chain from the 2 constituent chains. The order of linking of chains is prioritised such that the most likely linkages are made first. Structural variant calls from GRIDSS that are shown to be linked empirically by GRIDSS are linked first followed by a set of heuristics to prioritise the remaining uncertain links. This process continues to extend the length of and reduce the number of chains in each cluster until no further links can be made.

Constraints on linking breakends

In general each pair of facing breakends is considered by LINX as a candidate chain. However, 3 key constraints are applied in limiting candidate pairs:

- **'Trans' phased breakends are not permitted to chain -** A breakend may not be chained to another breakend within it's anchoring supporting distance unless it is linked by assembly.

- **Closed loops without centromeres are only allowed for ecDNA** - With the exception of ecDNA (see below), derivative chromosomes are assumed to connect to either a centromere or telomere at each end. Hence, chains which do not cross a centromere are not allowed to form closed loops (ie. the 2 breakends flanking a chain are not allowed to join to each other) except when specifically identified as ecDNA.

- **A chain should not pass through a region with lower available allele copy number than the JCN of the chain** - Chains are generally assumed to take place on a single paternal chromosome except in the rare case that 2 distinct overlapping paternal alleles of the same chromosome are involved . The copy number for the affected allele is calculated for each segment adjacent to each cluster by first determining the copy number of the unaffected allele across each contiguous set of breakends in the cluster and subtracting that from the total copy number. The total JCN of linked chains crossing a segment should not exceed the calculated allele copy number for that segment. In some noisy copy number regions or where both paternal alleles are involved we may not be able to determine an undisrupted allele and the allele copy number is not calculated. Additionally, if all possible links have been exhausted using this rule, the rule is relaxed such that chaining can continue under the assumption that allele specific copy number may not have been estimated accurately.

Uniform JCN clusters

Uncertainty in JCN measurement is one of the key confounding factors in predicting the derivative structure. Many clusters however do have the same JCN for all variants and these uniform JCN clusters are significantly simpler to chain into a derivative chromosome as each pair of breakends can only be linked with a single JCN.

Hence, each cluster is tested to see whether all its SVs could be explained by a single JCN value, using each SV's JCN estimate and range. If all SVs have a JCN range which covers the same value, even if not an integer, then the cluster is considered to be of uniform JCN and no replication will occur in the chaining routine. Furthermore, if all SVs have a JCN min less than 1 and a max > 0.5, then the cluster is also considered uniform.

Variable JCN clusters

For all clusters that cannot be resolved by a single uniform JCN, further considerations apply to explain the amplification of parts of the cluster. Biologically, each SV initially occurs by joining 2 breakends with a JCN of 1. However, a single chain in a cluster may contain SVs with different ploidies as a result of a replication from either a foldback inversion in a breakage fusion bridge event, or a later tandem duplication of part of the derivative chromosome. In these scenarios, the duplicated variants will appear multiple times in a single chain either repeated in the same direction in the case of a duplication or inverted in the case of a foldback.

Duplication events are permitted to ‘replicate’ a chain in 2 different ways:

1. **Foldbacks:** Foldbacks with half the JCN of another chain are permitted to link both their breakends to the same breakend of that chain, making a new chain of half the JCN with the other unconnected breakend of the non foldback chain at both ends of the new chain. In this foldback replication case the new chain may also be treated as a ‘chained foldback’ and extended further in the same manner if possible.

2. **Complex Duplications:** Conversely, complex duplications with half the JCN of another chain are permitted to join both of their breakends simultaneously to either ends of the other chain, effectively duplicating the entire chain, but keeping the same start and end breakends with half the JCN.

Partial replication of a chain is also possible by later variants which affect chains that have also been duplicated. In this case the higher JCN chain is split into 2 separate chains, one which is linked to and given the JCN of the lower JCN chain, and the other which is given the residual JCN.

Implementation of chaining algorithm

Linx first resolves all assembled and transitive SVs in the cluster into chains. LINX then keeps a cache of a set of chains consisting initially of all ‘single variant’ chains (ie. lone SVs) and assembled chains. Each chain has 2 breakends, a JCN and a JCN uncertainty. Each potentially linkable pair of facing breakends (subject to the available allele copy number and no closed loops rules described above) is also cached as potentially linked chains.

The following steps are then applied iteratively to join chains together until no more chain links can be made:

1. Apply priority rules below to choose the most likely linked pair of chains from the cache
2. Merge chains:
    a. Create new combined chain and calculate the JCN & JCN uncertainty
    b. For replication events, replicate and halve the JCN of the chain, inverting if it is a foldback type event.
    c. In the case of partially split chains, the higher JCN original chain is kept with its residual JCN
    d. Remove merged chains
3. Update cache of linked breakend pairs that include a breakend on the merged chains

Prioritisation of chain links

Where more than 1 possible pair of linkable chain exists in the cache, links are ranked and chosen by the following criteria in descending order of importance:

1. Links with available allele copy number*
2. Links containing highest JCN foldback or complex duplication chain
    a. Links where it splits another chain with 2x its JCN
    b. Links where it matches the JCN of another chain
    c. Links where it splits another foldback with greater than 2x JCN
    d. Links where it is itself split by another foldback or complex duplication with half the JCN
3. Breakends with a single link possibility
4. Links with highest matching JCN status (MATCHED > OVERLAPPING_JCN_RANGE > NO_OVERLAP)
5. Adjacent links
6. Higher JCN links (allowing for 0.5 abs and 15% threshold)
7. Shortest link

* For the purpose of the available allele copy number rule, a junction with an offsetting inferred breakend at the exact base is assumed to create a link if and only if the JCN of both the junction and the offsetting inferred breakend is greater than adjacent major allele copy number (allowing for 0.5 abs and 15% threshold)

For uniform JCN clusters, only rules 1, 3,5 & 7 are considered in the prioritisation of links.

A cluster can be chained into more than 1 chain, each one representing the neo-chromosomes resulting from the rearrangement event. Chaining is often imperfect and incomplete due to the inclusion of single breakends and uncertainty about JCN and subsequent breakend replication.

Special considerations for LINE clusters

LINE clusters typically involve one or more insertions from a single source location to multiple target sites in the genome, with occasional inversion or rearrangement of the inserted sequence. Due to the highly localised origin and frequent overlap of these inserted elements, the above chaining rules are not appropriate for LINE clusters. Instead, assembled and transitive links are chained first and then pairs of SVs or assembled/transitive chains that make consistent insertions to a single remote site are chained together. Single breakends at an insertion site that are paired with a breakpoint to a known or suspected LINE source region and have insert sequence alignment matching the same source LINE region with the correct orientation are also chained at the source site. SVs that cannot be paired off to form consistent insertions are not chained.

Special considerations for extrachromosomal DNA (ecDNA)

ecDNA (or double minutes) and the SVs contained within them are subject to special chaining rules in LINX. They are therefore identified prior to chaining. The key principle used to identify ecDNA is to look for high JCN junctions adjacent to low copy number regions which can be chained into a closed or predominantly closed loop.

LINX use the following algorithm to identify ecDNA

- Identify candidate ecDNA clusters: the cluster contains at least 1 non deletion junction with one breakend with a JCN > 5 AND at least 2.3x the adjacent major allele copy number. For clusters with maxJCN < 8, or if the only candidate DM variant is a single duplication, then the JCN of all candidate breakends must be at least 2.3x the adjacent major allele copy number
- Find all other potential ecDNA junctions in candidate clusters: any variant with JCN > max(5,25% of max ecDNA candidate JCN) and >2x the adjacent major allele copy number is classed as a candidate ecDNA junction.
- Attempt to chain candidate ecDNA junctions into closed segments: LINX tries to chain the identified candidate ecDNA junctions both with uniform and variable junction copy number and chooses the chain with the most closed segments
- Determine whether chained ecDNA junction meets ecDNA criteria: All of the following criteria must be met either for a closed loop or for the all the DM candidate variants as a whole if a closed loop cannot be formed:
    - At least one pair of breakends must be chained to form a closed segment (ignoring non overlapping deletion junctions)
    - Either a complete closed chain is formed or the number of closed breakends must be at least double the number of open breakends. If the max ecDNA candidate JCN < 8, then the entire chain MUST be closed.
    - The total length of closed segments must be > 1500 bases. If the cluster includes only closed segments enclosed by adjacent single or inferred breakends on both sides, then at least one closed segment must have Purple depthWindowCount > 5
    - The sum of the JCN from foldbacks in the cluster + the sum of JCN of junctions from regions internal to the ecDNA segment bounds to regions external (excluding assembled templated insertions) + the maximum JCN of any single or inferred breakend (excluding proximate pairs of clustered breakends with opposite orientation) on closed segments + 4 < reference JCN of DM
        - This rule is intended to ensure that the JCN of the DM could not have been achieved via amplification in a linear chromosome

        - If no foldbacks are DM candidate variants, then the reference JCN is set to the maximum JCN of any candidate DM SV in the cluster. Otherwise it is set to the lowest JCN of the foldback DM candidate variants. This is to allow for the uncertainty of whether foldbacks are part of the presumptive ecDNA or may have been part of a BFB event that caused the amplification in a linear chromosome

If LINX determines that an ecDNA event has occurred using the above criteria, it will retain and annotate the ecDNA chaining. Other junctions in the cluster with both breakends fully contained within a closed ecDNA segment are then only allowed to link to other variants within the ecDNA OR to the ecDNA forming variants. These links are likely lower JCN disruptions which occurred after the ecDNA was first replicated are present on a subset of the ecDNA.

Special considerations for clusters with 2 inversions or 2 translocations

Chains consisting of 2 inversions or 2 translocations that do not meet the criteria for ecDNA are common. Where the breakends overlap they may have multiple plausible paths, in each case either forming a reciprocal inversion / translocations or a deletion / duplication with templated insertion. In cases where breakends cannot be phased, LINX can not uniquely distinguish between these 2 scenarios. LINX implements the following heuristics to attempt predict the event type:

For a pair of translocations:

- If the breakends face away from each other or are cis-phased on both chromosomes, then resolve as RECIP_TRANS (no chaining)
- If the breakends face towards each other and are not cis-phased on both chromosomes then:
    - If double minute rules are satisfied then resolve as DOUBLE_MINUTE
    - If the segment on either chromosome is bounded by LOH resolve then chain the other segments and resolve as a DUP_TI chain.
    - Else resolve as RECIP_TRANS_DUP (no chaining)
- If the breakends face towards each other and are not cis-phased on one chromosome and the breakends face away from each other or are cis-phased on the ohter chromosome then:
    - If the segment on the chromosome with facing breakends is bounded by LOH resolve then chain the other segments and resolve as a DEL_TI chain.
    - Else resolve as RECIP_TRANS_DEL (no chaining)

Similarly, for clusters with a pair of inversions with opposite orientations:

- If the 2 inversions face away from each other and overlap on their outer breakends only, then resolve and chain as a RECIP_INV
- If the 2 inversion face towards each other and overlap on their inner breakends only, then:
    - If double minute rules are satisfied resolve as DOUBLE_MINUTE
    - If either segment is bounded by LOH, chain the other segment and resolve as DUP_TI chain
    - Else chain the outer breakends and resolve as RECIP_INV_DUP
- If one inversion fully encloses the other then:
    - If both inversions have the same JCN and the shorter possible templated insertion is bounded by LOH then chain the shorter templated insertion and resolve as a DEL_TI
    - Else if the inner inversion length < 100K, chain the shorter templated insertion and resolve as RESOLVED_FOLDBACK
    - Otherwise chain the longer of the 2 templated insertions and resolve as a RECIP_INV_DEL_DUP
- If there are 2 foldback inversions facing each other without overlap then
    - If double minute rules are satisfied then resolve as DOUBLE_MINUTE
    - Else resolve as FB_INV_PAIR (This may be formed from a breakage fusion bridge event)

Chain annotations

The following data is captured for each templated insertion in a chain:

- Whether the link is assembled
- Distance to the next link and whether it traverses any other breakends or links
- Genic overlap and any exact exon boundary exon matches (ie. a pseudogene)

Annotation of local topology

Consecutive breakends with no more than 5kb between them or which are part of the same foldback inversion are grouped together into a local topology group and given an id. The number of TIs formed by chained segments wholly within the local topology group are counted and a topology type is given to the remaining variants based on the breakend orientations. The topology types are categorised as one of the following (after excluding all TIs)

- TI_ONLY - All breakends in the group form templated insertions
- ISOLATED_BE - One breakend only
- DSB - A par of breakends forming a deletion bridge
- FOLDBACK - One foldback only
- FOLDBACK_DSB - A foldback with the outer breakend forming a deletion bridge
- SIMPLE_DUP - A single DUP of <5k bases
- COMPLEX_LINE - Any other cluster that is resolved as LINE
- COMPLEX_FOLDBACK - Any other cluster that includes a foldback
- COMPLEX_OTHER - Any other cluster

4. Gene impact & fusion prediction

Annotation of breakends with potential gene impact

For each breakend we search for genes that could be potentially disrupted or fused by the structural variant. To do this we find and annotate the breakend for any transcript that either:

- Has an exon or intron overlapping the breakend
- Has its 5' end downstream of and facing the breakend and less than 100k bases where no other splice acceptor exists closer to the breakend.

Each breakend is additionally annotated for the transcript with the following information:

- **disruptive**: a breakend is disruptive for a particular transcript if the SV is an inversion, translocation or single breakend or a deletion/duplication that overlaps at least part of an exon in the transcript AND the variant is NOT part of a chain which does not disrupt the exon ordering in the transcript. A breakend which is resolved as type 'LINE insertion' is never marked as disruptive.
- **transcript coding context**: UPSTREAM, 5_UTR, CODING, 3_UTR, DOWNSTREAM OR NON_CODING
- **gene orientation**: relative orientation of gene compared to breakend (UPSTREAM or DOWNSTREAM)
- **exonic** (TRUE/FALSE)
- **exact base phase**: The exact base phasing of the current location
- **Next splice site information**: The distance to, phasing of and exon rank of the 1st base of the next facing splice acceptor or donor (note: phasing will be different from exactBasePhase if the breakend is exonic in the transcript or the coding context is upstream. Null if there are no subsequent splice sites in the gene.
- **exon total count**: The total number of exons in the transcript (for reference)
- **transcript biotype**: The ensembl biotype of the transcript

Known pathogenic fusions and promiscuous partners

Configuration of pathogenic fusions impacts LINX in 2 ways:

1. Some criteria for fusion calling are relaxed for known fusions due to the high prior likelihood of pathogenic fusions
2. As well as attempting to predict all fusion events, LINX uses the configured list of fusions to determine a subset which are 'reported' as likely pathogenic.

To produce the list of known fusions provided with LINX, a broad literature search was performed to find a comprehensive list of well-known fusions that are highly likely to be pathogenic. The criteria used for inclusion of a particular fusion in the curated list was either multiple independent reports of the fusion, or single case reports with either convincing demonstration of the pathogenicity in a model system or clear response to a therapy targeted to the specific fusion.

The curated fusions were classified into 3 categories:

- **Known fusions** (n= 396) – these are transcript fusions which fuse either the coding regions of 2 genes to form a novel protein or the 5' UTR regions of 2 genes which may lead to increased expression of the 3' partner. A well-known example is TMPRSS2_ERG
- **Known IG enhancer rearrangements** (n= 17) – these are structural rearrangements in B-Cell lymphomas and leukemias that relocate enhancers from one of the @IG regions (IGH,IGK,IGL) to increase expression of a 3' partner. A well-known example is @IGH-MYC
- **Known exon deletions & duplications** (n=11) – these are deletions or duplications of exons in specific exon ranges of a handful of genes which are known or highly likely to be pathogenic. Common examples are EGFR vII and vIII

A set of 'promiscuous' fusion partners was also determined from this list so that potential novel fusions with fusion partners that have been identified in multiple fusions previously can also be reported as potentially pathogenic. Any gene which was identified in 3 or more known fusions was marked as a promiscuous 5' partner and likewise if it was identified in 3 or more known fusions in our curated list was marked as a promiscuous 3' partner. MYC and CRLF4 were also marked as 3' promiscuous since they feature in known fusions with both IG enhancer and known fusions. FGFR1 is also added as a 5' promiscuous partner.

For 12 promiscuous fusion partners {FGFR1, FGFR2, FGFR3, TMPRSS2, SLC45A3, HMGA2,BRAF, RET, ALK, ROS1, ETV1, ETV4} a specific exon range has been identified as highly promiscuous and is identified in the knowledge base. Fusion reporting criteria are relaxed in these ranges. 10 promiscuous 3' genes {BRAF ,RET ,ROS1 ,ALK ,MET ,NRG1 , NRTK1, NTRK2 & NTRK3) are marked as 'high impact' and also have special treatment in the fusion reporting logic.

A section of @IGH gene stretching from the diversity region to the end of the constant region was also marked as a 'promiscuous IG partner' as it features in many IG fusions.

Fusion Prediction

Identify fusion candidates

Fusions are predicted in LINX by looking for consecutive and novel splice donor-acceptor pairings that are joined together in derivative chromosomes by either a single structural variant or a continuous chain of structural variants.

For each single SV and for every facing pair of SVs in the same chain identify all viable splice acceptor and splice donor fusion combinations which satisfy the following conditions:

- 3' gene partner must have coding bases
- 5' gene partner transcript must have one of the following ensembl biotypes: 'protein_coding','retained_intron','processed_transcript','nonsense_mediated_decay','lincRNA'
- The upstream breakend must fall within the 5' partner transcript and be disruptive to the transcript.
    - The downstream breakend must fall either within the 5' gene or within 100kb upstream.
    - The combined length of all segments in the chain must be less than 150kb.
- The SV or chain must join appropriate contexts of the 5' and 3' genes (see table below) and for coding regions must be inframe after allowing for any skipped exons. For exonic to exonic fusions exact base phasing is also checked as splice acceptor to splice donor phasing. The following table shows allowed contexts:

	3’ Fusion Partner Context					
5’ Partner Context	Upstream	5’UTR Intronic	5’UTR Exonic	Coding Intronic	Coding Exonic	3’ UTR or Non-Coding
Upstream	X (Enhancer)	X (promoter loss)				X (No Downstream Impact)
5’UTR or non-coding Intronic	YES (1)	YES	YES	alt start codon (4)		
5’UTR or non coding Exonic	exon skipped (3)		YES	exon skipped & alt start codon (3,4)	alt start codon (4)	
Coding Intronic	YES (1)	SEE NOTE (5)		YES	YES (2)	
Coding Exonic	exon skipped (3)				YES	
3’UTR	X (post stop codon in upstream gene)					

(1) for breakends in the upstream region of the 3’ partner, the 1st exon is not considered as it does not have a splice acceptor and so the 2nd exon is assumed to be the 3’ fusion partner. 5’ partner coding to 3’ partner upstream is also possible if the 3’ partner coding region starts in the 1st exon.
(2) If fusing intron to exon, the fusion occurs with the next downstream exon, so check against the frame of the end of the exon instead of the exact base.
(3) Exonic to Intronic can occur if alternative splicing causes exon with exonic breakend to be skipped
(4) 5’ partner 5’UTR or non-coding to coding region of 3’ partner can technically make a fusion, but would need to find 1st alternative start codon also in-frame. These are called as out of frame and only reported for known fusions
(5) Coding Intronic to non-coding allowed only when transcript starts on 1st base of the next downstream exon - in this case we fuse to the first base of the gene which is allowed.

Special rules for single breakends which align to a site which forms a pathogenic fusion

The post processing steps of GRIDSS annotate the best alignments for the insert sequence of single breakends which cannot be uniquely mapped. If any single breakend has an alignment which would create fusion matching a known fusion in our knowledge base then that fusion is called as if the single breakend was a translocation to that alignment.

Special rules for IG rearrangements

In the special case of IG enhancer rearrangements, the rearrangement normally occurs either between the ‘D’ and ‘J’ region (due to RAG mediation D-J recombination failure – common in IGH-BCL2 fusions) or in the switch region just upstream of the constant regions (due to failure of isoform switching mechanisms – common in IGH-MYC rearrangements). In the former case, the Eμ enhancer is the likely driver of elevated expression whereas in the latter the driver is likely the alpha 1,2 & 3 regulatory region enhancer. To predict a relevant rearrangement, LINX requires that the breakend in IG is oriented downstream towards the enhancer regions and is connected to the 5’ UTR of the 3’ gene partner.

Special gene specific cases

LINX also has special rules to support unusual biology for a handful of known pathogenic fusions:

- For CIC_DUX4 and IGH_DUX4, the DUX4 end may map to a number of different chromosomal regions, including the telomeric ends of both chromosomes 10q and 4q and the hg19 alt contig GL000228.1.

- For RP11-356O9.1_ETV1 fusion (pathogenic in Prostate cancer) the breakend on the 5’ side breakend is permitted to be up to 20kb downstream of RP11-356O9.1.
- In the case of IGH-BCL2, LINX also looks for fusions in the 3’UTR region and up to 40k bases downstream of BCL2 facing in the upstream orientation towards BCL2 (common in Folicular Lymphomas).

Prioritise genes and transcripts

Each candidate chained splice acceptor and splice donor fusion pair may have multiple gene fusion transcripts on both the 5’ gene and 3’ gene that meet the above criteria. Occasionally genes may also share a splice acceptor or splice donor in which case the transcripts of that groups of genes are considered together. LINX prioritizes the potential transcript candidates and choose a single pair of 5’ and 3’ transcripts via the following criteria in order of priority:

- Fusion is a KNOWN_PAIR or known EXON_DEL_DUP
- A phased fusion is possible
- Chain is not terminated early
- 3’ partner biotype is ‘protein_coding’
- 3’ partner region is INTRONIC or EXONIC
- No exons are skipped
- Best 3’ partner transcript, ranked by canonical and then longest (non NMD) protein coding
- Best 5’ partner transcript ranked by canonical, then longest protein coding, then longest gene

If multiple chains link the same 2 genes then they are prioritised again according to the above logic.

Reportable fusions

In addition to predicting fusions, LINX also tries to identify likely viable pathogenic fusions and marks as reportable. To maximise precision whilst ensuring high impact fusions are always likely to be reported, the criteria vary by fusion type with more relaxed criteria for known pathogenic pairs due to high prior likelihood. High impact promiscuous fusion partners which may be clinically relevant (including NTRK1-3, BRAF, RET, ROS1, ALK) also have more relaxed criteria

The criteria are summarised in the below table.

Criteria	KNOWN PAIR	IG KNOWN PAIR	EXON DEL DUP	HIGH IMPACT PROMISCUOUS	PROMISCUOUS OTHER
Knowledge-base match	GENE PAIR	GENE PAIR	Breakends within EXON RANGE	INTERGENIC ONLY	INTERGENIC ONLY
‘Nonsense Mediated Decay’ biotype allowed for 3’ partner	FALSE	FALSE	FALSE	FALSE	FALSE
Maximum chain links	4	4	4	4	4
Maximum upstream distance for 3’ partner	100kb	10kb	NA	100kb	10kb
Phasing	INFRAME or SKIPPED EXONS	NA	INFRAME or SKIPPED EXONS**,***	INFRAME or SKIPPED EXONS	INFRAME****
Allow early chain termination or disruption by intermediate splice acceptor or donor	TRUE	TRUE	TRUE	FALSE	FALSE

* 5’UTR to coding regions are also allowed for KNOWN_PAIR fusions (>1Mb length)
** The breakend and the fused exon must both be in the specified ranges on both 5’ and 3’ side
*** Out of frame also reported for exonic to exonic EXON_DEL_DUP only (under assumption of possible phased indel)
**** Skipped exons are allowed if a known exon range is configured, but only if the breakend and fused exon must be within the specified range on the promiscuous side, and any skipping is allowed on the non-promiscuous gene.

Additionally, LINX checks that the protein domains retained in the 4' partner may form a viable protein. Specifically The following domains must be preserved intact in the 3' partner if they exist: Ets domain; Protein kinase domain; Epidermal growth factor-like domain; Ankyrin repeat-containing domain, Basic-leucine zipper domain,High mobility group box domain. The Raf-like Ras-binding domain must be disrupted if it exists (mainly affects BRAF).

Finally, LINX sets a likelihood for each reported fusion. KNOWN_PAIR, IG_KNOWN_PAIR and EXON_DEL_DUP are set to HIGH likelihood. PROMISCUOUS fusions are set to HIGH likelihood only if the fused exon matches the known exon range, or else LOW otherwise.

Amplification, deletion and disruption drivers

Homozygous disruption drivers

LINX can optionally take as input a catalog of point mutation, amplification and homozygous deletion drivers which is created by PURPLE based on the raw somatic variant data and determined copy number profile. LINX leverages it's chaining logic to extend the driver catalog by searching for 2 additional types of biallelic disruptions which disrupt all copies of the gene but do not cause a homozygous deletion in an exonic segment (which is PURPLE's criteria for homozygous deletion). Specifically LINX searches for 2 additional types of homozygous disruptions:

- **Disruptive breakends -** Any pair of disruptive breakends that form a deletion bridge or are oriented away from each other and both cause the copy number to drop to <0.5 after allowing for the JCN of any overlapping deletion bridges.
- **Disruptive duplications -** Any duplication which has both breakends disruptive in the transcript and a JCN >= flanking copy number at both ends.

Linkage of drivers to contributing structural variant clusters

We link each driver in the catalog that is affected by genomic rearrangements (ie. high level amplifications, homozygous deletions and biallelic point mutations in TSG with LOH and the homozygous disruptions found by LINX) to each structural variant cluster which contributed to the driver. 1 or more structural variant clusters may contribute to each event and/or the driver may be caused by a whole chromosome or whole arm event which cannot be mapped to a specific variant but which has caused significant copy number gain or loss

Amplifications drivers are linked to either one or more clusters which explain the gain in copy number over the gene (marked as type 'GAIN'), a gain in copy number in the centromere (marked as type 'GAIN_ARM') or across the whole chromosome (marked as type 'GAIN_CHR'). More than 1 of these factors can be recorded as a cause of amplification if its copy number gain is at least 33% of the largest contributing factor. To determine whether a cluster contributes to gene amplification, the copy number change of all its breakends surrounding the gene are summed into a net cluster copy number gain, and the copy number loss of any opposing clusters are subtracted. If a net gain remains, then the cluster is considered as contributing to the gene amplification.

Homozygous deletion drivers are linked to clusters that are either directly bound by the homozygous deleted region (svDriverType = DEL) or a cluster that bounds the LOH (svDriverType = LOH). Biallelic point mutations drivers in TSG with LOH are linked only to the cluster that bounds the LOH (svDriverType = LOH). If the LOH bounds extend to the whole arm or whole chromosome, then a record is created with linked clusterId is set to NULL and the svDrivertype is marked as LOH_ARM or LOH_CHR respectively.

The supported linkages between drivers and SVs are summarised in the table below

Driver Type	Events per driver	svDriverTypes
Amplification	1+	GAIN GAIN_ARM GAIN_CHR
		DEL

Homozygous Deletion	2	LOH LOH_ARM LOH_CHR
Biallelic point mutation in TSG	1	LOH LOH_ARM LOH_CHR

LINX visualisation

LINX provides functionality to present detailed visualisation of genomic rearrangements including genic impact in CIRCOS[7] format. LINX writes a set of ‘VIS' files in a specific format which form the base data to generate the visualisation. The visualisation tool only depends on these files and so in principle any tool could provide SV & CN data in this format.

There are 3 main components in the output of the LINX visualisation each described below.

CIRCOS view

The CIRCOS view shows either the complete set of rearrangements in a cluster or set of clusters if 1 or more cluster ids are specified (‘cluster mode’) or all clusters that touch a chromosome if a chromosome is specified (‘chromosome mode’).

The CIRCOS view has 6 tracks showing from innermost to outermost:

1. Junctions
2. Minor allele copy number profile
3. Copy number profile
4. Derivative chromosomes
5. Impacted genes (if any)
6. Affected chromosomes

The scaling of both distances and copy numbers in the figure has been modified to make the figure readable. Specifically, the distances between each feature in the chart (either breakend or gene exon start or end) are modified to a log based scale, so that the entire genomic rearrangement spanning millions of bases and multiple chromosomes can be viewed, but that local topology of regions with high densities of breakpoints can be introspected. JCN is set with a linear scale but is scaled down if the maximum cluster JCN exceeds 6 or if the total density of events exceeds a certain maximum to ensure that even the most complex clusters can be introspected. Additionally, if the total number of breakends displayed exceeds XX then the junctions become increasingly transparent such that other features on the plot don’t become obscured.

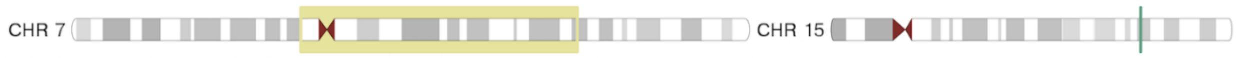
Another key feature of the CIRCOS plot is the ability to trace the derivative chromosome(s). Each segment in the 4th track represents a segment of the derivative chromosome and is linked on both ends either to a centromeric or telomeric end (marked with an open or closed square respective) or a breakend (marked with a track or triangle in the case of foldbacks). Each derivative chromosome can be traced continuously from one telomeric / centromeric end to another (or to a single breakend if one is reached) by following a continuous series of segments and breakends. To make this easier to follow, each time a new segment is connected on a chromosome the segment is offset outwards slightly. Hence the derivative chromosomes can be traced from the inside to the outside of the 4th track of the diagram. A cluster may contain 1 or more derivative chromosomes. In cluster mode each derivative chromosome will be shown in a different colour for ease of viewing (with a maximum of 10 colours after which all derivative chromosomes are shown in black). In chromosome mode, derivative chromosomes will be shown in the same colours, but each cluster is shown in a different colour. Red and green are reserved for simple deletions and tandem duplications respectively. Since there may be many of these on a single chromosome, telomeric and centromeric connectors are not shown for these simple variant types. Light blue is also reserved for LINE clusters which can also be frequent in samples with highly deregulated LINE machinery.

Most of the possible annotations are shown in the LINX visualisation guide (see figure 1B)]. Additionally, 2 types of genomic regions which are frequently disrupted in tumor genomes are indicated using light grey shading on the copy number regions in the 3rd track. For known LINE source regions the green copy number section is shaded and for known fragile sites the red copy number section is shaded light grey.

Chromosome view

Since the CIRCOS only represents a part of the genome, the chromosome view is provided to indicate which parts of the genome is shown. Each of the chromosomes included in the cluster(s) shown is displayed. The part of the chromosome that is included in the figure is highlighted in the colour matching the colour used in the outer ring of the CIRCOS. The banding and location of the centromere on each chromosome is also shown.

An example of the chromosome view is shown below indicating the cluster includes a large section of chromosome 7 including the centromere and a small slither of chromosome 15 on the Q arm:

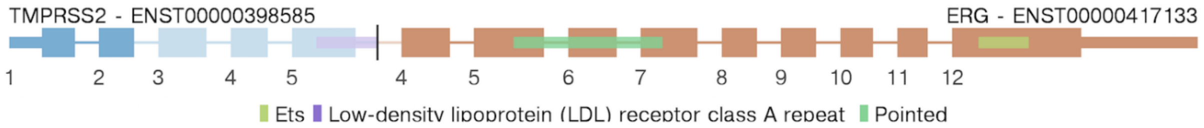


Fusion view

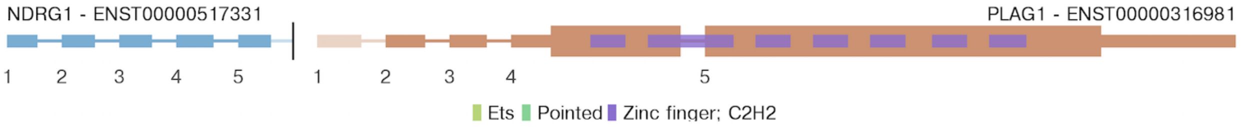
The fusion view is added for reportable fusions only in LINX. It's purpose is to show the predicted structure of the fused gene. The fusion includes the fused segments of both the 5' and 3' partner in blue and red and always reads from left to right. The gene representation for each genes and follows the standard conventions of thick bands for coding regions, thinner bands for 5' UTR and 3' UTR exonic regions and thin lines to represent the intronic sections. Protein domains are shown in coloured bands across the exons which they include and are labeled in the accompanying legend. As with the CIRCOS view the lengths of exonic gene segments in the fusion view are scaled by a log scale to improve readability. The intronic segments are set to a fixed segment length regardless of the length.

The fused gene is shown up to and including the breakend on either side that is connected either directly in the case of a simple fusion or via a chain in a chained fusion. If LINX predicts that one or more exons are skipped, then the skipped exonic segments and protein domain sections are faded.

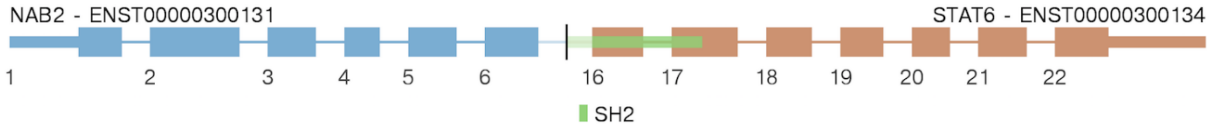
For example in the following TMPRSS2-ERG fusion, exons 3, 4 & 5 are faded and the LDL domain is also faded, indicating that LINX predicts these exons are skipped in order to make a viable in-frame protein, despite the break end occuring after the 5th exon:


TMPRSS2 - ENST00000398585
ERG - ENST00000417133
1
2
3
4
5
4
5
6
7
8
9
10
11
12
Ets
Low-density lipoprotein (LDL) receptor class A repeat
Pointed

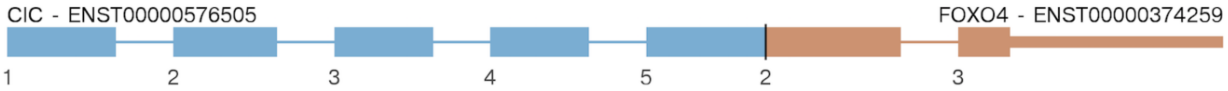
A similar case can be seen in this example where the 5'UTR region of NDRG1 is fused upstream of the 1st exon in PLAG1. Since the 1st exon of a gene has no splice acceptor, LINX predicts the 1st exon is skipped and it is faded on the chart with the fusion connecting to the start of exon 2 which also begins in the 5'UTR region of PLAG1:



Even when no exons are skipped, a protein domain may be partially disrupted if it extends to an exon the other side of the breakpoint. In the below example, you can see the SH2 protein domain extends before exon 16 and is faded and disrupted:



Whilst fusions are normally intronic, rare exonic to exonic fusions do occur. The below figure shows a CIC-FOX04 example where the 2 exons are directly fused



Examples of LINX visualisations on COLO829T

COLO829 is a widely studied melanoma tumor-normal paired cell line that is frequently used as a somatic reference standard for benchmarking in next generation sequencing. This note uses COLO829 LINX output as an example to give insight into how to interpret a diversity of visualisations in tumor genomes.

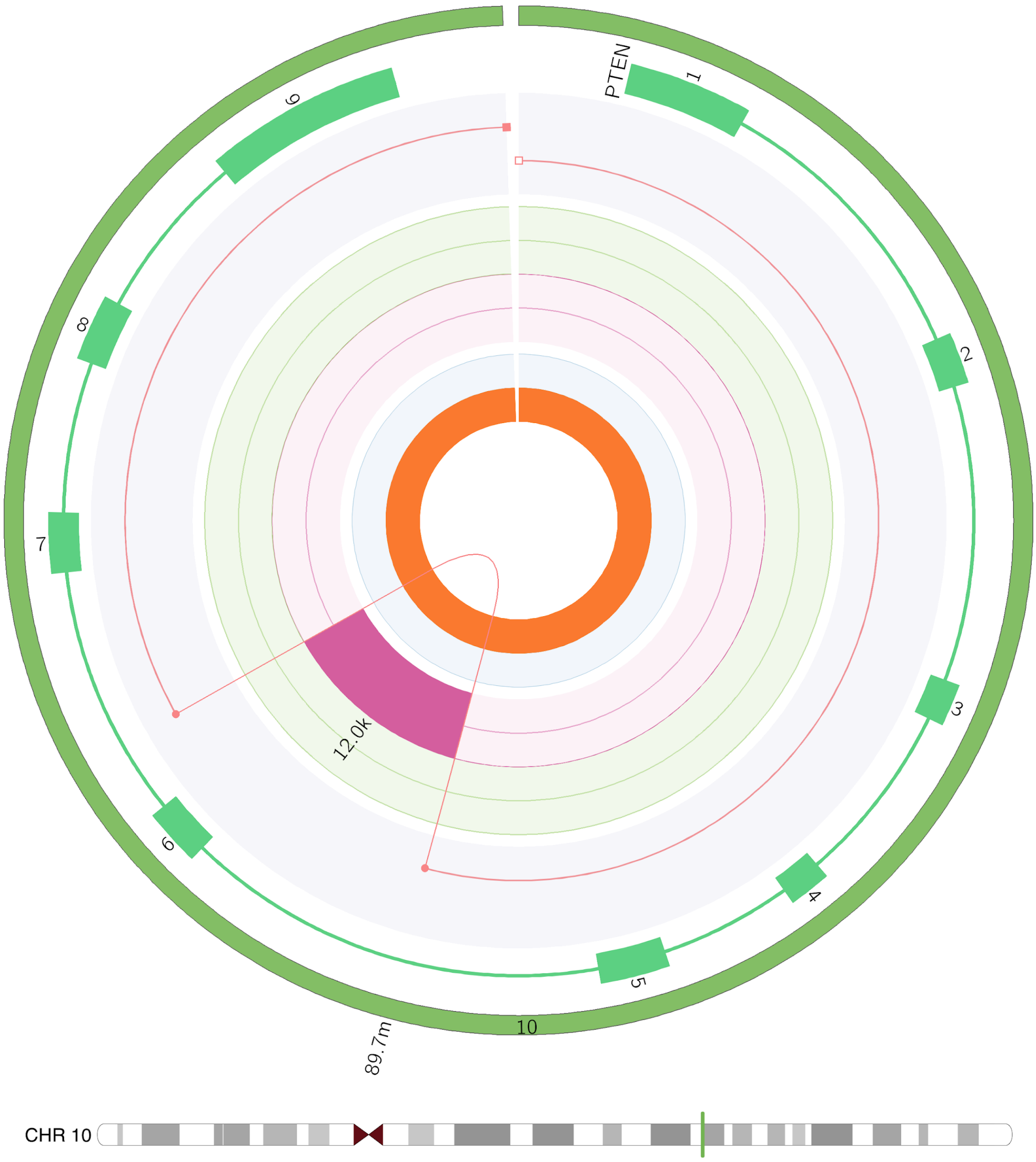
LINX identifies 61 clusters in COLO829T. One cluster (Cluster 0) is a low VAF translocation which is filtered as 'ARTIFACT' by LINX. The remaining 60 clusters consist of

- 9 simple tandem duplications
- 35 simple deletions
- 2 synthetic deletions
- 1 synthetic unbalanced translocation
- 1 pair of single breakends that is treated as an implied duplication
- 2 LINE insertions
- 6 complex clusters
- 4 incomplete clusters (all unclustered single or inferred breakends)

Example plots for each of the types of simple clusters and all of the complex classified by LINX in COLO829T are described below. The visualization guide in figure 1B of the manuscript can be used to aid interpretation of the figures.

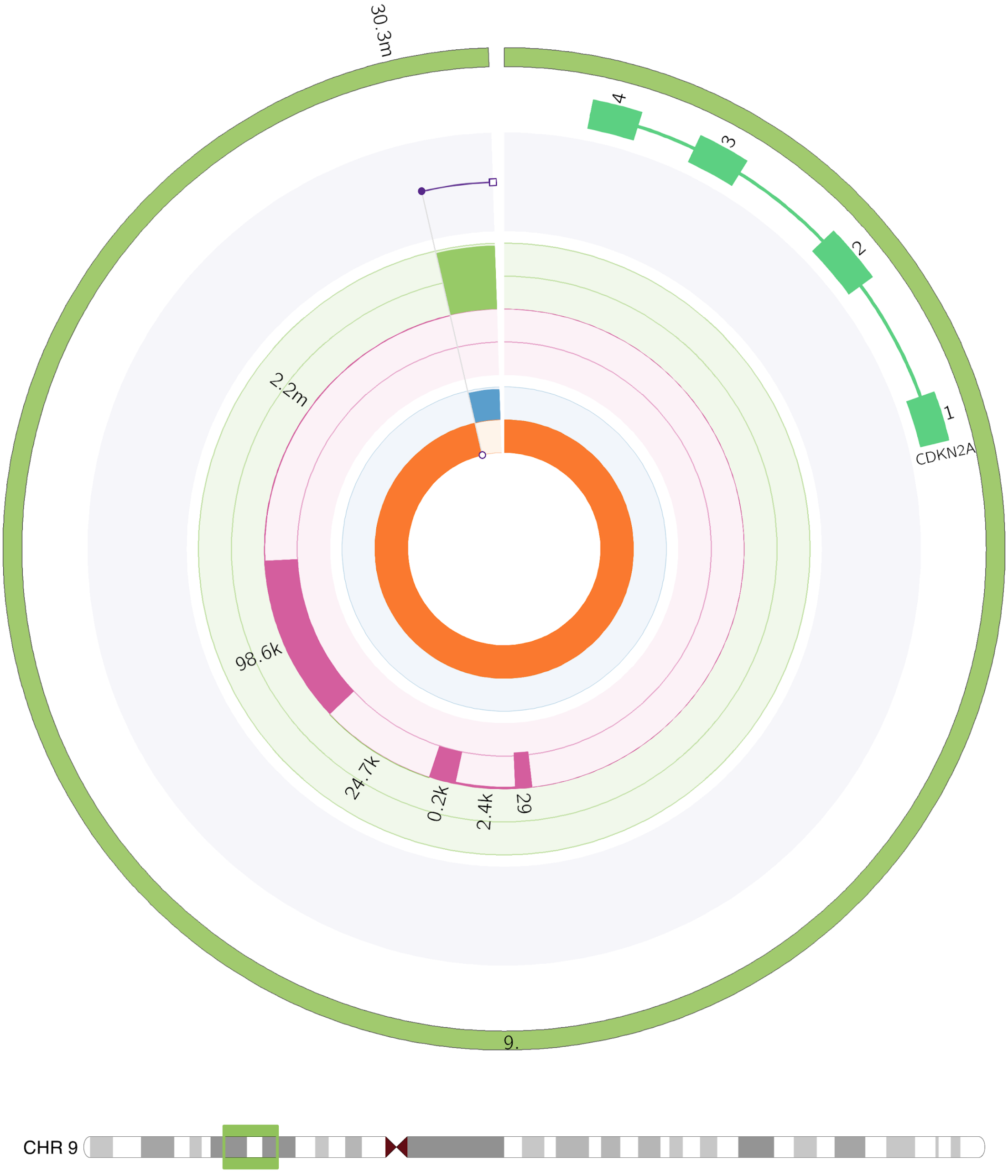
Simple deletion

Cluster 67 is a simple deletion on chromosome 10 of 12kb that leads to a homozygous loss of exon 6 in PTEN. Note that simple deletions and tandem duplications will not be drawn by LINX by default unless they affect a gene.



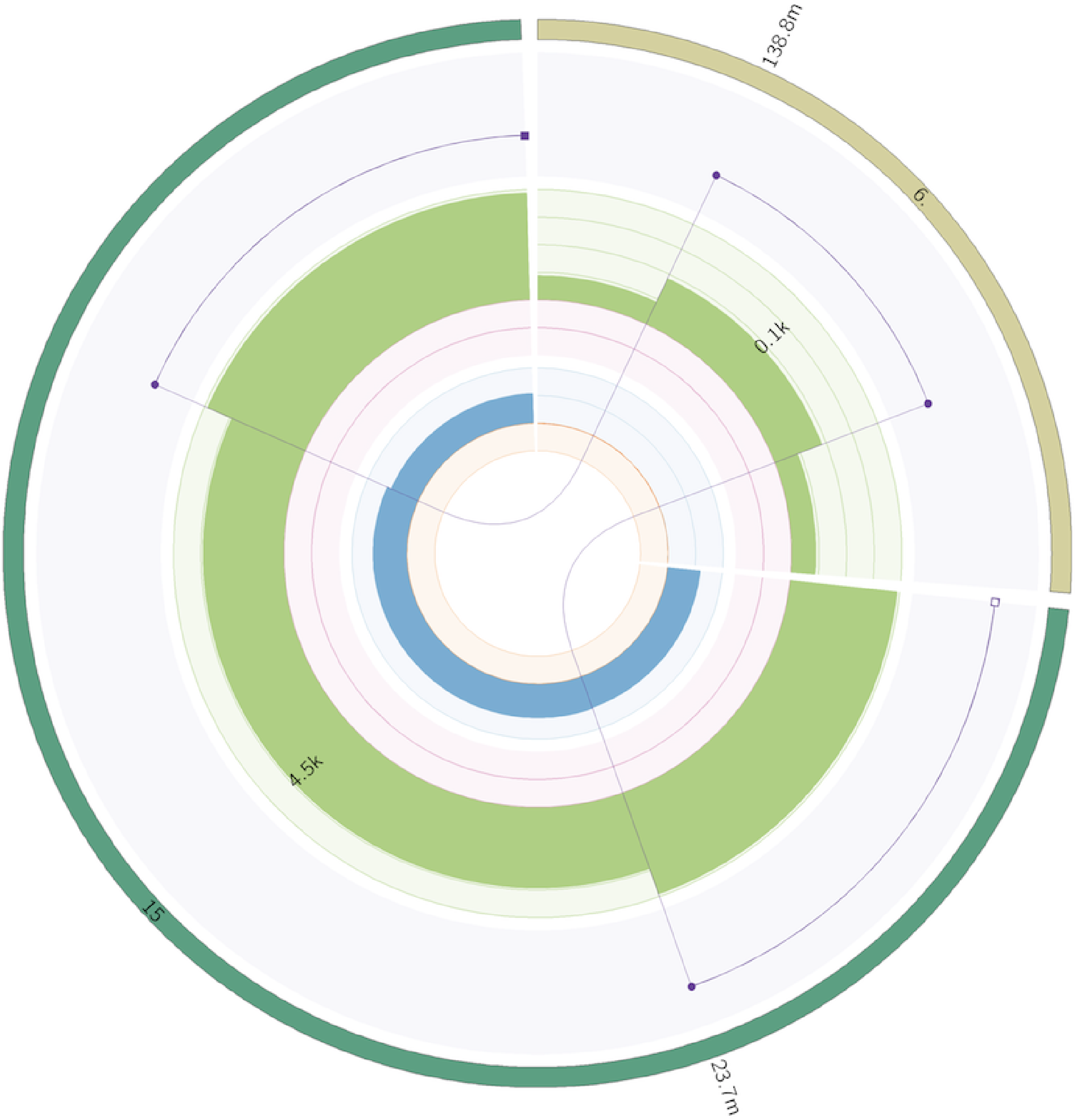
Single breakend (incomplete cluster)

Cluster 59 is a single breakend on chromosome 9 with JCN of 3, that causes a LOH from 30.3M to the 9P arm telomere. The other end of the single breakend is unmappable. The LOH contributes to a biallelic driver in CDKN2A which also has a missense variant (not shown)



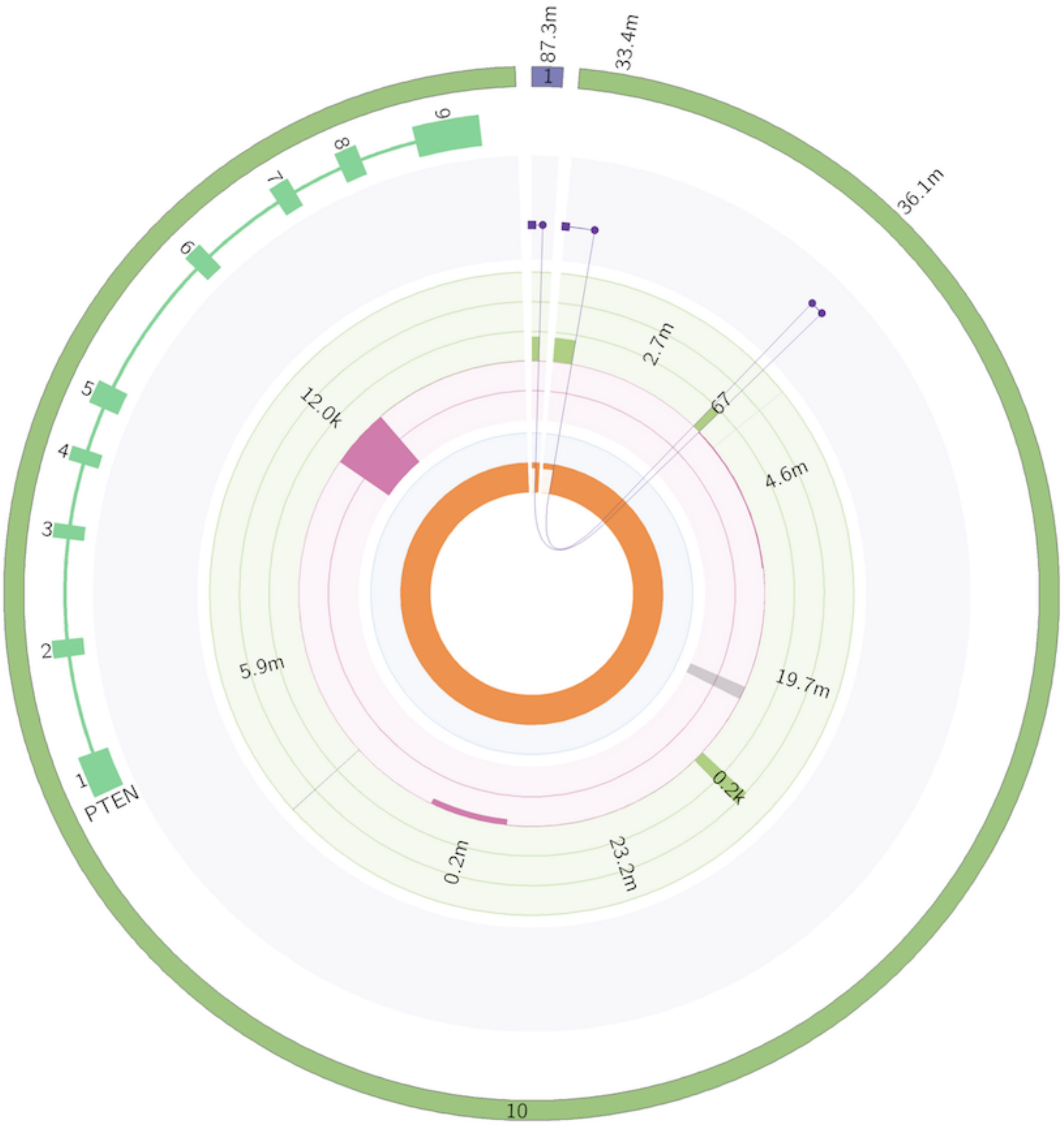
Synthetic deletion

LINX simplifies variants that form shards of <1kb insertions and will classify as ‘synthetic’ if removal of the shards allows it to be resolved as a 1 or 2 break cluster type. Cluster 13 from COLO829T is an example of a synthetic deletion on chromosome 15 with a 100 base insertion from an otherwise undamaged region of chromosome 6



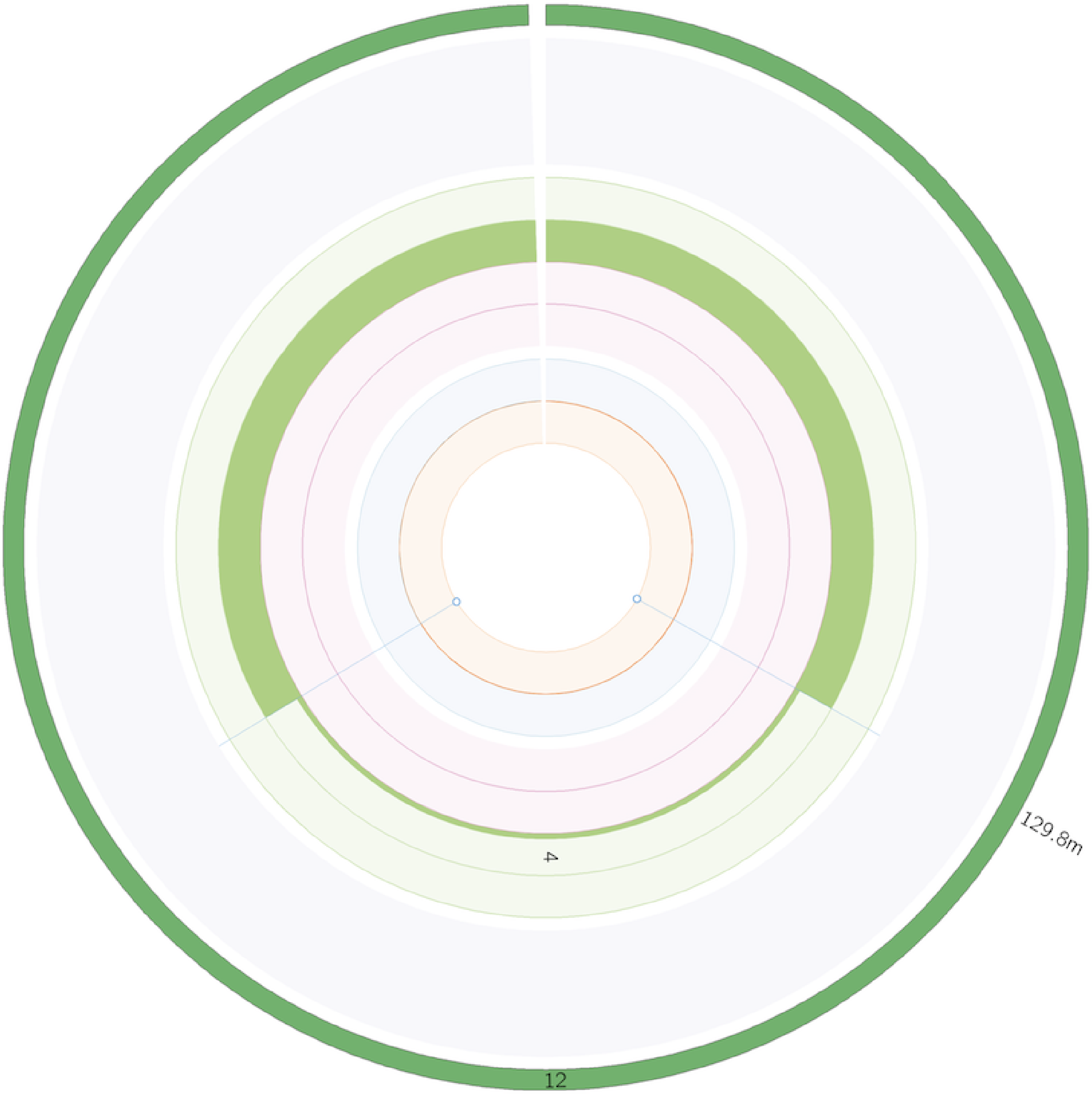
Synthetic unbalanced translocation

Cluster 64 is an unbalanced translocation has occurred between chromosome 1 and 10 causing a LOH for much of on chromosome 1 and chromosome 10 including the PTEN driver gene (the copy number impact of a homozygous deletion of exon 6 of PTEN can also be seen on the circos in pink, but from an unrelated 2nd hit event). A small shard of 67 bases has been inserted from chromosome 10 in the unbalanced translocation, so the event is classified as a ‘synthetic’ unbalanced translocation.



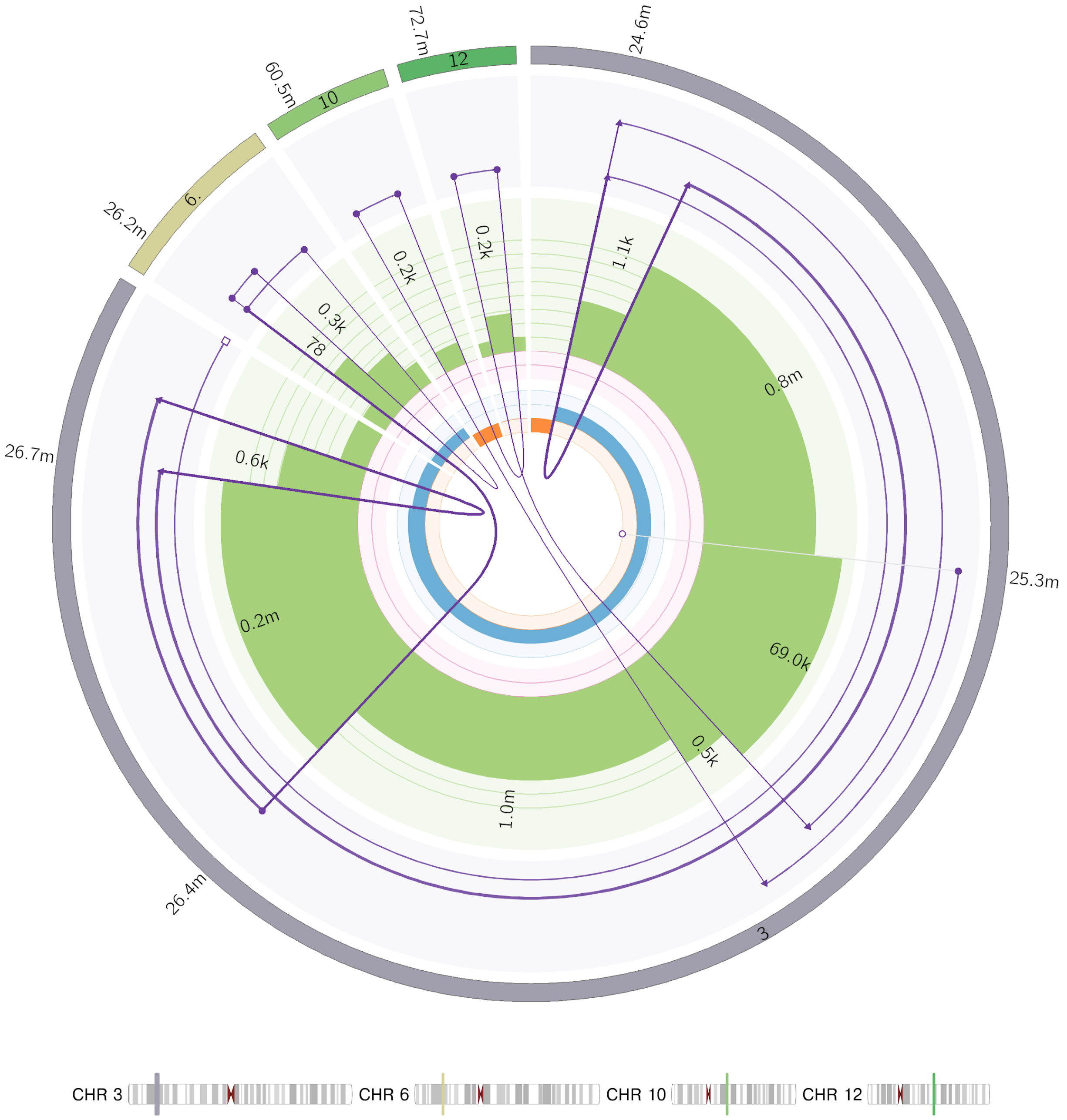
LINE insertion

2 single breakends form the insertion site of a likely LINE insertion on chromosome 12. The LINE source element cannot be uniquely mapped in this case and the LINE classification is made based on the POLYA insertion sequence (not shown). The open circles on the edge of the innermost circle are used by LINX to represent single breakends.

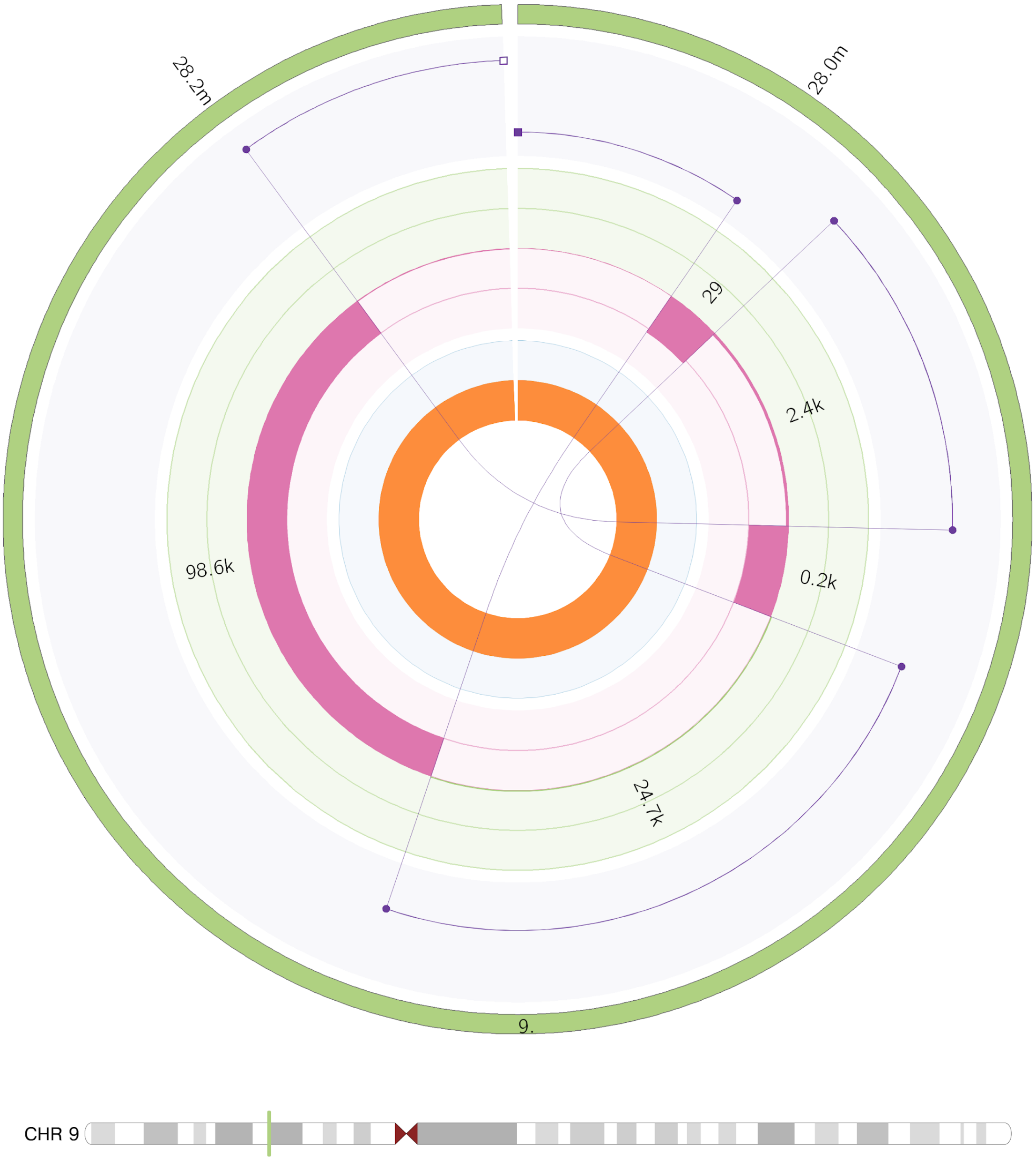


COMPLEX clusters

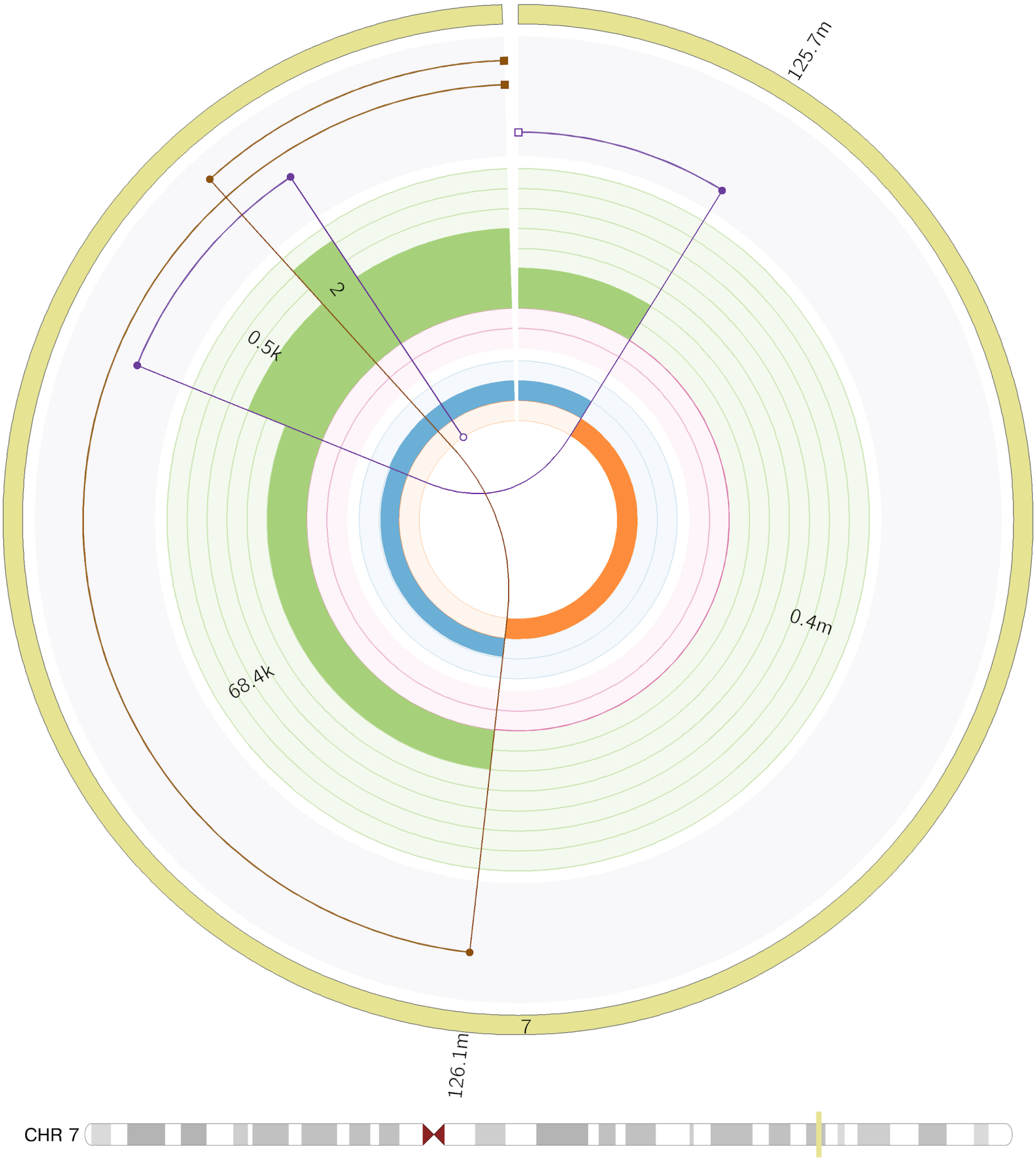
6 clusters are classified as COMPLEX by LINX in COLO829T. The most complex rearrangement is a highly amplified structure predominantly on chromosome 3 that includes shards inserted from chromosome 6, 10 and 12 without further DNA damage on those chromosomes. The amplification is caused by 2 simple foldback inversions and 2 chained foldback inversions (with the inserted shards), with foldback breakends identified with a triangle marker in the image. A single breakend at position 25.3M is also clustered and likely resolves the breakage fusion bridge cycle:



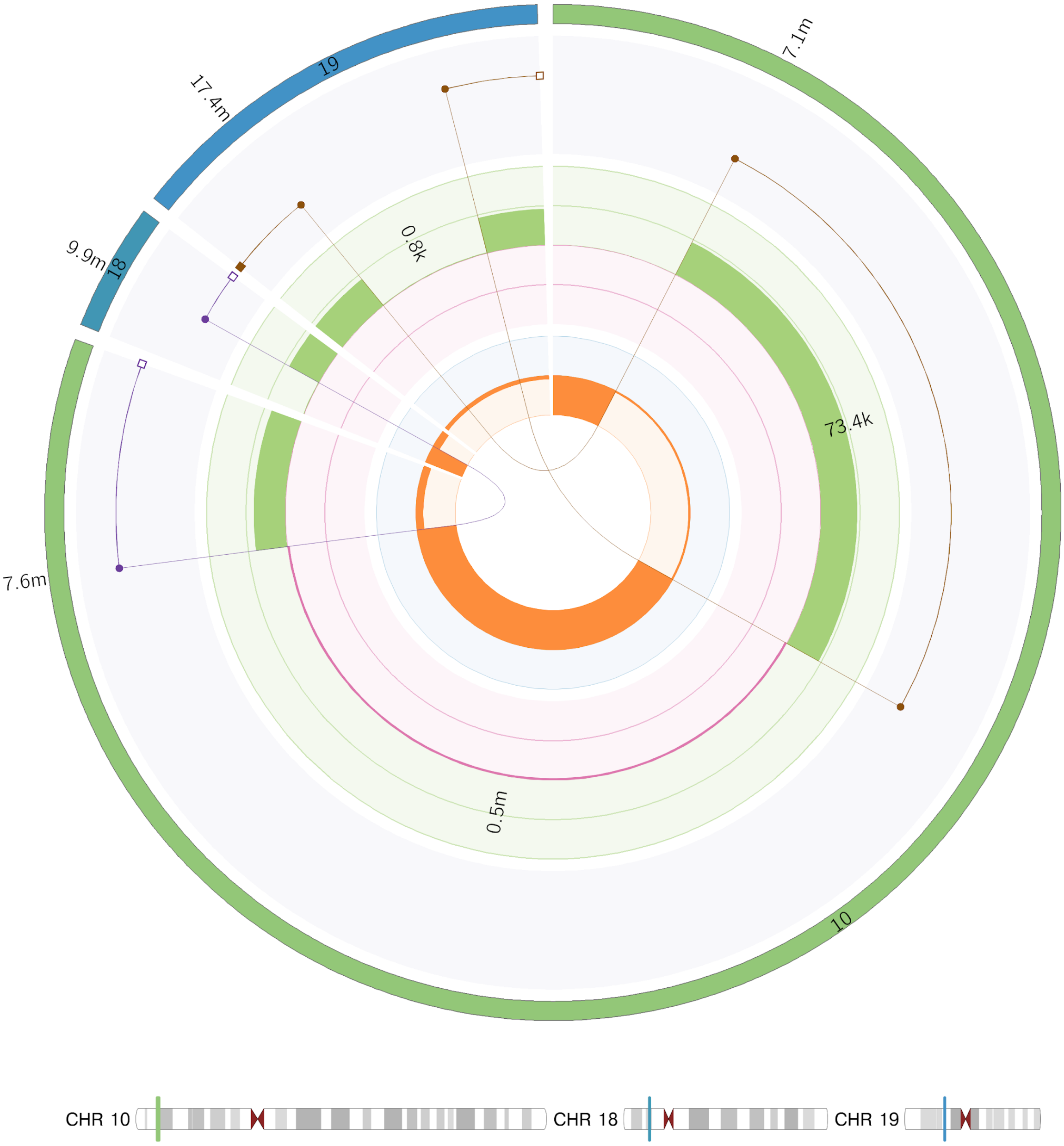
A complex event also occurred on chromosome 9 between 28M and 28.2M and involves 2 inversions and a duplication, likely caused by multiple concurrent double stranded breaks in this region 2 segments are retained and 1 long segment and 2 short gaps are lost.



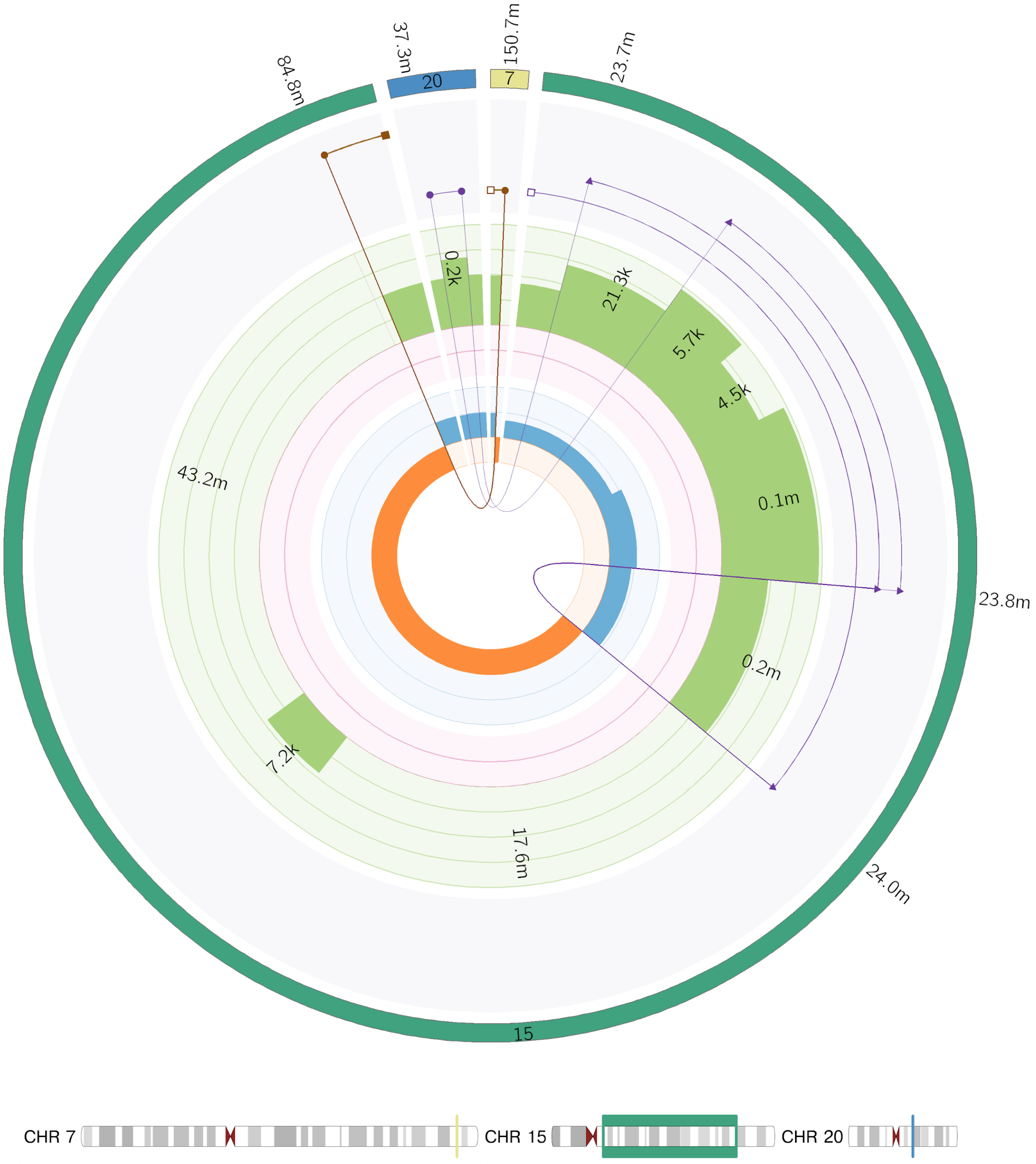
A 3rd complex rearrangement is on chromosome 7 and involves an inversion, a duplication and a single breakend. The rearrangement has caused a LOH in the region from 125.7M to 126.1M and amplification at the telomeric end of the chromosome 7Q arm. It is unknown where the single breakend connects to in the genome



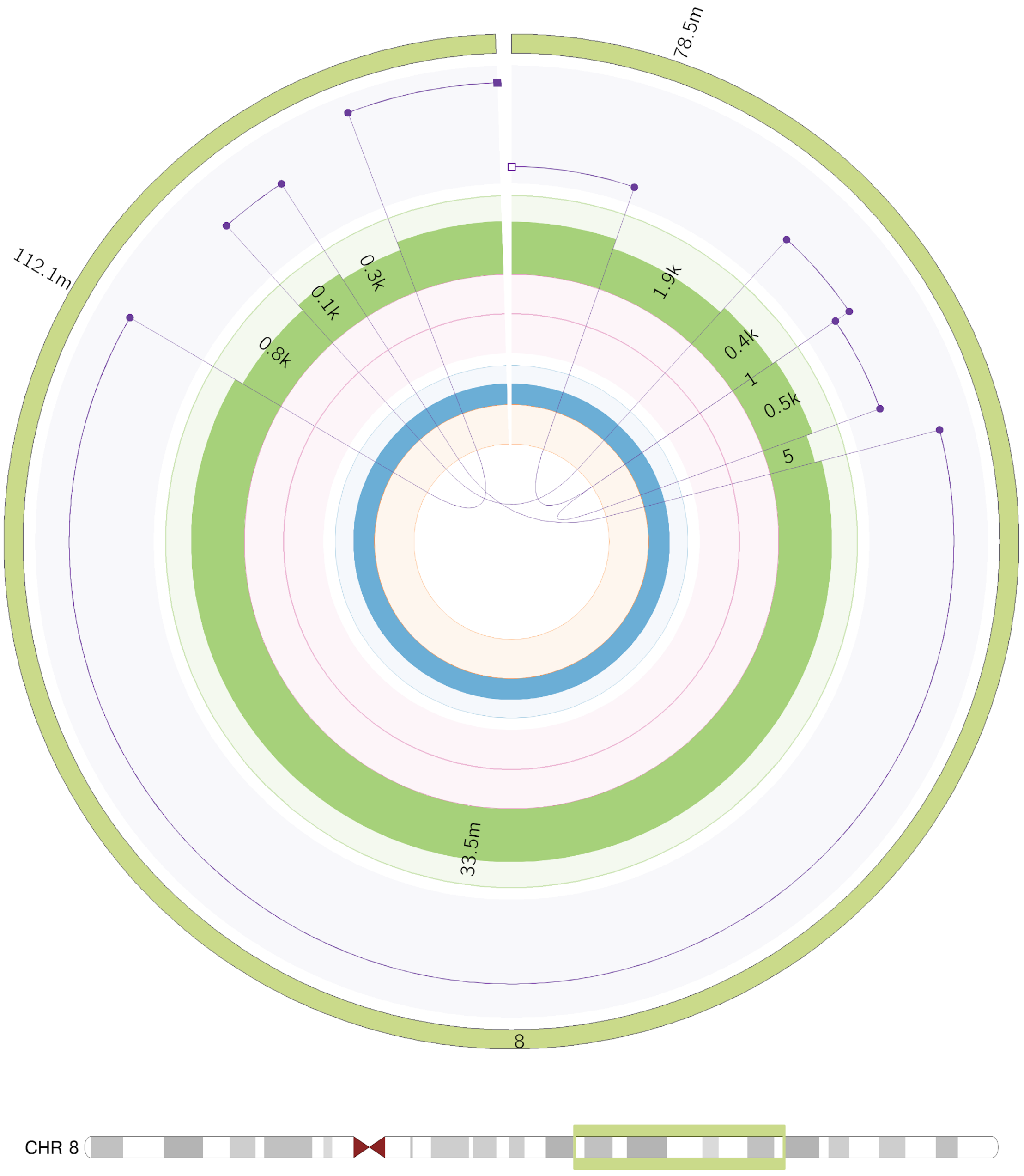
The 4th complex rearrangement involves 3 translocations on 2 distinct derivative chromosomes. The first derivative chromosome inserts a 73.4kb segment of chromosome 10 into a 800 base gap in chromosome 19 and is shown in brown. The 2nd derivative chromosome (purple) is a translocation from chromosome 10 to 18 which has caused amplification of the centromeric end of the 18P arm. The 2 derivative chromosomes' events are clustered together due to the LOH deletion bridge in between. An LOH of the first 7.1M bases of chr 10 has also occurred in the same event.



A 5th complex rearrangement primarily involves chromosome 15, causing a LOH on a large segment of chromosome 15 from 24m to 84.8m. The q telomere of chromosome 15 is linked by translocation to the q centromere of chromosome 7. A pair of facing foldbacks amplify the centromeric region of 15Q. One of the foldbacks is synthetic with a short shard of 200 bases inserted from chromosome 20.

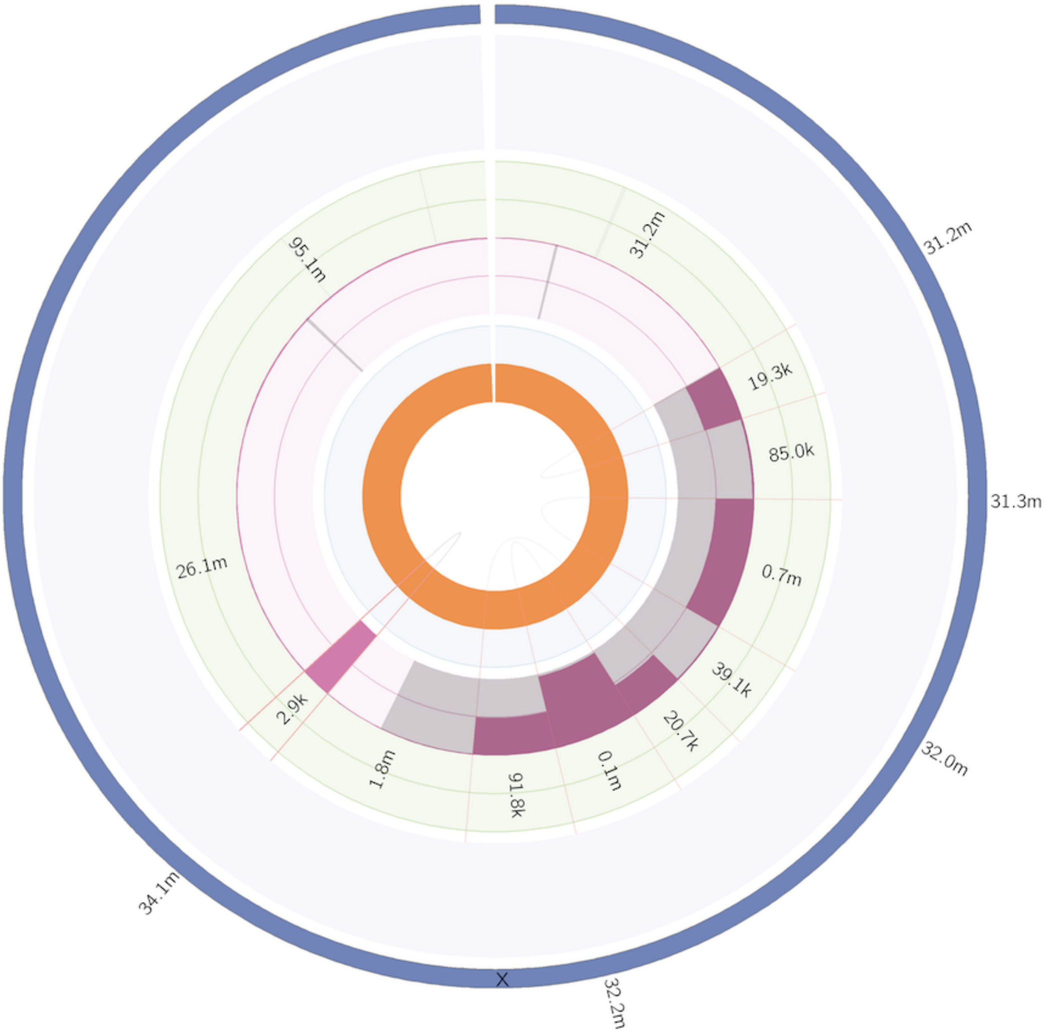


The final complex rearrangement is a subclonal rearrangement of 5 junctions on chromosome 8 that affects 2 regions of the Q arm of chromosome 18 around 78.5M and 112.1M. At both loci multiple breaks have occurred leading to 3 short chromosomal segments and 1 long segment being rearranged. The damage appears to be localised as the derivative chromosome can be chained consistently from centromere to telomere.



Chromosome view

An alternative way to view LINX output is by chromosome view. This may be used to visualise all the clusters with breakends originating from a particular chromosome. The below example shows multiple deletions on Chromosome X. Note, for simple deletions, duplications and insertions, the chromosomal segments are not shown as there can be many in highly rearranged samples. Only the junction is shown with a fixed red colour for deletions, green for duplications and light blue for insertions. Other clusters will each have a distinct colour. Note that for known LINE source elements the green copy number section is shaded and for known fragile sites the red copy number section is shaded light grey. In COLO829, 4 of the 5 deletions on chromosome X are contained within the DMD fragile site as shown.



References

1. Li, Y., Roberts, N.D., Wala, J.A., Shapira, O., Schumacher, S.E., Kumar, K., Khurana, E., Waszak, S., Korbel, J.O., Haber, J.E., et al. (2020). Patterns of somatic structural variation in human cancer genomes. Nature *578*, 112–121.

2. Bignell, G.R., Santarius, T., Pole, J.C.M., Butler, A.P., Perry, J., Pleasance, E., Greenman, C., Menzies, A., Taylor, S., Edkins, S., et al. (2007). Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. Genome Res. *17*, 1296–1303.

3. Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M., et al. (2013). Punctuated evolution of prostate cancer genomes. Cell *153*, 666–677.

4. Rodriguez-Martin, B., Alvarez, E.G., Baez-Ortega, A., Zamora, J., Supek, F., Demeulemeester, J., Santamarina, M., Ju, Y.S., Temes, J., Garcia-Souto, D., et al. (2020). Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nat. Genet. *52*, 306–319.

5. Priestley, P., Baber, J., Lolkema, M.P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., et al. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. Nature *575*, 210–216.

6. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

7. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.