Decoil: Reconstructing extrachromosomal DNA structural heterogeneity from long-read sequencing data

Mădălina Giurgiu[1,2,3,4], Nadine Wittstruck[1,2,3], Elias Rodriguez-Fos[1,2,3], Rocío Chamorro González[1,2,3], Lotte Brückner[1,2,3,5], Annabell Krienelke-Szymansky[1,2,3], Konstantin Helmsauer[1,2,3], Anne Hartebrodt[6], Richard P. Koche[7], Kerstin Haase[1,2,3†], Knut Reinert[4†], Anton G. Henssen[1,2,3,5†]

[1]Department of Pediatric Oncology and Hematology, Charité – Universitätsmedizin Berlin, Berlin, Germany.
[2]Experimental and Clinical Research Center of the Max Delbrück Center and Charité Berlin, Berlin, Germany.
[3]Charité–Universitätsmedizin Berlin, Berlin, Germany.
[4]Freie Universität Berlin, Berlin, Germany.
[5]Max Delbrück Center for Molecular Medicine, Berlin, Germany.
[6]Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany.
[7]Center for Epigenetics Research, Memorial Sloan Kettering Cancer Center, New York, USA.

Contributing authors: madalina.giurgiu@charite.de; anton.henssen@charite.de;
†These authors contributed equally to this work.

Abstract

Circular extrachromosomal DNA (ecDNA) is a form of oncogene amplification found across cancer types and associated with poor outcome in patients. ecDNA can be structurally complex and contain rearranged DNA sequences derived from multiple chromosome locations. As the structure of ecDNA can impact oncogene regulation and may indicate mechanisms of its formation, disentangling it at high resolution from sequencing data is essential. Even though methods have been developed to identify and reconstruct ecDNA in cancer genome sequencing,

it remains challenging to resolve complex ecDNA structures, in particular amplicons with shared genomic footprints. We here introduce Decoil, a computational method which combines a breakpoint-graph approach with *LASSO* regression to reconstruct complex ecDNA and deconvolve co-occurring ecDNA elements with overlapping genomic footprints from long-read nanopore sequencing. Decoil outperforms *de-novo* assembly methods in simulated long-read sequencing data for both, simple and complex ecDNAs. Applying Decoil on whole genome sequencing data uncovered different ecDNA topologies and explored ecDNA structure heterogeneity in neuroblastoma tumors and cell lines, indicating that this method may improve ecDNA structural analyzes in cancer.

Keywords: long-read, ecDNA, nanopore, reconstruction, heterogeneity

1 Introduction

Extrachromosomal DNA (ecDNA) is an important form of oncogene amplification in cancer [1], which can be formed through multiple mechanisms [2–4]. As a result, ecDNA can be structurally diverse, with different functional outcomes. The structure of ecDNA can impact gene regulation through the rearrangement of regulatory elements as well as topologically associated domain (TAD) boundaries [5]. To explore ecDNA diversity and complexity, high-resolution computational methods to reconstruct ecDNA with high accuracy from genome sequencing data are required. The reconstruction of ecDNA from sequencing data remains challenging due to the variable complexity and intratumor heterogeneity of these elements. On the one hand, a single ecDNA can be heavily rearranged and contain low-complexity sequence regions (e.g. repeats), which pose a challenge to mapping and *de-novo* assembly based methods. On the other hand, one tumor can contain different ecDNA elements [6, 7], which can either originate from different or shared genomic locations [8]. The latter scenario may be very challenging for ecDNA reconstruction, as different co-occurring ecDNA elements have overlapping genomic footprints, making it difficult to attribute the overlapping features to each of the different circular elements. In the past years, several computational tools have been developed to reconstruct ecDNA from different input data. Some methods were developed to detect circularized DNA regions by identifying the breakpoints leading to circularization (circle-enrich-filter [9], Circle-Map [10], ecc_finder [11]). These approaches are suitable for detecting simple circular amplicons, but overlook complex ecDNA structures. To overcome these limitations, more recently, methods focused on reconstructing complex ecDNA based on different technologies, e.g. short-read whole-genome sequencing [12], optical-mapping combined with short-read sequencing [13], and long-read sequencing were developed [14]. Lastly, methods have been developed to delineate ecDNA structural heterogeneity [6], by isolating and reconstructing individual ecDNA elements, leveraging *a priori* knowledge about the ecDNA present in the sample of interest. However, a method that reconstructs complex ecDNA structures and captures heterogeneity by distinguishing between ecDNA elements with overlapping genomic footprints from whole-genome sequencing data

without such *a priori* knowledge is still largely missing to date. We here present Decoil, a computational method to reconstruct genome-wide complex ecDNA elements and deconvolve ecDNAs with shared genomic sequences from bulk whole-genome long-read sequencing using Nanopore technology. Decoil is a graph-based approach integrating the structural variant (SV) and coverage profiles to discover and reconstruct complex ecDNAs. It uses *LASSO* regression to infer likely ecDNA structures and estimate their proportions, by accounting for overlapping genomic footprints. This may improve future studies of ecDNA structural heterogeneity.

2 Results

2.1 An overview of the Decoil algorithm

Decoil reconstructs complex ecDNA structures from long-read nanopore sequencing data using aligned sequencing reads, structural variants and coverage profiles as input (Figure 1a). The genome is initially fragmented using a clean breakpoints set (Figure 1a #1). A weighted undirected multigraph is build to encode the structural rearrangements, where nodes are defined as genomic non-overlapping segments and edges represented the structural variants (Figure 1a #2). Next, the graph is explored using a depth-first search approach to discover genome-wide simple circular paths (Figure 1a #3). These can represent a unique circular element or be a sub-component of a more complex circular structure (Figure 1b). Subsequently, to account for circular elements containing nested circles, simple circular paths with at least one overlapping genomic fragment are merged into a derived larger circular structure. In order to identify the likely ecDNA elements present in the sample, all simple and derived circle candidates are leveraged as features to fit a *LASSO* regression against the read-alignment mean coverage profile. This model will (1) select the likely circles explaining the amplification and (2) estimate their proportions within the sample (Figure 1a #4). Using this approach, Decoil can account for ecDNA structures with overlapping genomic footprints Figure 1c). Lastly, a filtered confident set of circular paths is generated (Figure 1a #5), together with the annotated topology (as defined below), proportion estimates and reconstruction thread visualization (Figure 1a (#6+#7)).

2.2 Ranking and simulating ecDNA topologies to capture ecDNA structure diversity

Currently, no guidelines exist for the assessment of ecDNA reconstruction performance from long-read data, nor do benchmarks exist like those for single nucleotide variant (SNV), insertion-deletion (INDEL) and structural variant (SV) detection [15, 16]. The lack of a gold standard datasets for assessing ecDNA reconstruction makes the evaluation of Decoil contingent on high-quality simulated data. The read-alignment of an individual ecDNA generates a structural variant collection. This information was used as the basis to systematically rank the computational complexity of ecDNA topologies (Figure 1b). This provides an approach for performance evaluation based on modeling different SV's composition on the amplicon, i.e. deletions (DEL), duplications (DUP), inversions (INV), translocations (TRA) and inverted-duplications (INVDUP).

We propose seven ecDNA topologies (Figure 2): i. Simple circularization, ii. Simple SV's, iii. Mixed SV's, iv. Multi-region, v. Multi-chromosomal, vi. Duplications and vii. Foldbacks. These ecDNA topologies were leveraged to simulate rearrangements on the amplicon in order to create a representative and comprehensive collection of more than 2000 ecDNA templates (Figure 2a), based on which we generated *in-silico* long-read reads at different depth of coverage. This collection serves as a benchmark dataset for evaluating Decoil's reconstruction performance across varying computational complexities and could be a useful dataset for future ecDNA genomic studies.

2.3 Decoil's performance evaluation to reconstruct ecDNA in simulated data

The accuracy of ecDNA reconstructions was quantified using the normalized largest contig as a score to measure the assembly contiguity (Section 4.6). Decoil reconstructed simple ecDNA topologies with high-fidelity (largest contig normalized of 0.99 for more than 500 simulations) from simulated data, i.e. topologies i, ii, iii, iv and v (Figure 2c,d). For the complex topologies, i.e. vi and vii, Decoil reconstructed at least 60% of the true structure (largest contig normalized > 0.6, Figure 2d, Suppl. Table S2), in more than 70% of the simulations (total of > 1200 simulations). Poorly resolved structures (largest contig normalized < 0.6) often contained mixed rearrangements including nested duplications and foldbacks, suggesting that such ecDNA elements are more challenging to reconstruct computationally. To demonstrate the utility and feasibility of the method, we compared Decoil against Shasta *de-novo* assembler [17] on the simulated dataset, using different Quast metrics (e.g. largest contig, longest alignment, N50). For 70% of simple structures Shasta and Decoil largest contig covered at least 90% and 99% of the true structure, whereas for 70% of complex topologies only 30% and 60% were covered, respectively (Figure 2d). Decoil outperformed Shasta for both, simple and complex topologies in terms of structure completeness (Suppl. Table S1). Thus, Decoil enables the accurate reconstruction of simple and complex ecDNA to a greater extent as current state-of-the-art algorithms used for long read sequencing.

2.4 Decoil recapitulates ecDNA complexity and their co-occurrence in well characterized cancer cell lines

To show the versatility of the Decoil algorithm, we applied it to shallow whole-genome nanopore sequencing of three neuroblastoma cell lines, i.e. CHP212, STA-NB-10DM and TR14, for which ecDNAs were previously reconstructed based on various circular DNA enrichment methods and/or validated using fluorescence in situ hybridization (FISH)[5, 18, 19]. Decoil's reconstructions recapitulated the previously validated ecDNA element in CHP212 with high fidelity (Suppl. Fig. S1a,b). An ecDNA harboring *MYCN* and a gene fusion between *SMC6* and *FAM49A* was previously observed in STA-NB-10DM cells [18], which was confirmed by Decoil's reconstructions (Figure 3a). The ecDNA element in STA-NB-10DM also contained additional genes and was predicted to be 2.1 MB in size with an estimated 171 amplicon copies, harboring an interspersed duplication according to Decoil's reconstruction

(Figure 3a). Multiple co-occurring ecDNA elements, referred to as ecDNA species in a previous report, were observed in TR14 cells [19]. The three different ecDNA elements, containing *MYCN*, *ODC1* and *MDM2* were reconstructed by Decoil with high fidelity in TR14 (Figure 3b). Additionally, Decoil identified a previously unreported 1.09 MB (Suppl. Table S3) multi-chromosomal ecDNA element containing fragments from chromosome 1 and 2, with an estimated 22 amplicon copies, harboring *SMC6* and *GEN1* (Figure 3b). This is the largest amplicon and has the lowest number of estimated copies relative to the other co-occurring ecDNA elements, which may be the reason why other reports have not been able to identify it so far.

For comparison, the reconstruction's contiguity was evaluated in cell lines using the *de-novo* assembler Shasta. For CHP212, the agreement between Decoil and the Shasta was 100% (Suppl. Fig. S1b,c). In STA-NB-10DM, the interspersed duplication on ecDNA indicates increasing structural complexity and is more challenging to reconstruct. Thus, Shasta did not assemble a contiguous circular element (Suppl. Fig. S2a), whereas Decoil identified a contiguous circular path through the graph of this ecDNA element (Figure 3a). For TR14, the structures of amplicons harboring *SMC6*, *MDM2* or *ODC1* were consistent between Decoil and Shasta (Suppl. Fig. S3, Suppl. Fig. S2b). The *MYCN*-containing ecDNA was reconstructed by Decoil (Figure 2b), but was not fully resolved by Shasta (Suppl. Fig. S4b) due to overlapping rearrangements at the *MYCN* locus. Thus, Decoil is a versatile algorithm to (1) reconstruct complex ecDNA elements in cancer cell lines and (2) discover previously unknown ecDNAs from long-read sequencing data.

2.5 Decoil can recover ecDNA structure heterogeneity

To demonstrate that Decoil captures ecDNA heterogeneity, i.e. resolve structurally distinct ecDNA elements with overlapping genomic footprint, we generated 33 *in-silico* mixtures, by pair-wise combination of three neuroblastoma cell lines at different ratios, i.e. CHP212, STA-NB-10DM and TR14, each containing a structurally distinct ecDNA element with sequence overlaps at the *MYCN* gene (Figure 3d, Section 4.7). The individual amplicons were recovered in the different mixtures with an overall 93% amplicon breakpoint recall, which increased with the dilution fraction (Figure 3c). These results were dependent on the coverage and SV calling. Thus, Decoil can distinguish between different co-occurring ecDNA elements, even when they share similar sequences, enabling the measurement of structural ecDNA heterogeneity.

2.6 Exploring structural ecDNA complexity in cancer patients using Decoil

In order to explore structural ecDNA complexity in tumors, shallow whole-genome nanopore sequencing on a cohort of 13 neuroblastomas was performed, of which 10 harbored at least one ecDNA element as determined by FISH and three did not harbor ecDNAs and served as negative controls. One ecDNA-containing sample was removed from the analysis due to failed QC. Decoil did not detect any ecDNA in the negative control cohort and reconstructed at least one amplicon for the other 9 samples. The

reconstructed ecDNA elements varied greatly in their complexity (Figure 4f) and ranged from very simple (Figure 4a) or multi-region (Figure 4b) to heavily rearranged multi-fragmented structures (Figure 4c,d). Decoil reconstructed two ecDNA elements with an individual estimated proportions of more than 700x in patient #4, resolving the same breakpoints as previously described in single cell ecDNA sequencing data from this tumor ([7]). For samples with a very high structural-variant density at the genomic site of ecDNA origin, Decoil reconstructed multiple circular elements with different estimated relative proportions, which indicates ecDNA structural heterogeneity (Figure 4e). The reconstructed ecDNAs originated from chromosome 2 or chromosome 12. Multi-region topology, i.e. ecDNA originating from a fragments of the same chromosome, seemed to be the most frequent ecDNA topology identified in patients, consistent with the ecDNA elements detected in cell lines (Figure 4f). No Multi-chromosomal topology was detected in this cohort.

Decoil reconstructed ecDNA elements with a mean size of 1.4 MB in cell lines and 0.7 MB in patient samples (Figure 4g), which is in line with mean ecDNA sizes in other tumor sequencing studies [20]. Contiguous genomic fragments on ecDNA had a mean size of 127 kb in cell lines and 145 kb in patient samples (Suppl. Fig. S5b). While the ecDNA size was conserved for the different topologies (Suppl. Fig. S5a), complex ecDNA structures had significantly shorter fragments than simple ecDNA (Figure 4h, Suppl. Fig S5c). Lastly, simple ecDNA had higher copy numbers than complex ones in this cohort (Figure 4i, Suppl. Fig. S5d), in line with previous reports in neuroblastoma. This indicates that yet unknown structural features may influence ecDNA maintenance and/or oncogene regulation, resulting in differences in accumulation of ecDNA elements in large cancer cell populations.

3 Discussion

The structural complexity and heterogeneity of ecDNA make its reconstruction from sequencing data a challenging computational problem. We here presented Decoil, a method to reconstruct co-occurring complex ecDNA elements.

Due to their random mitotic segregation, many ecDNA elements, which may structurally differ, co-occur in the same cancer cells [7]. Disentangling ecDNA with shared genomic regions has not yet been addressed by other methods, and it cannot be resolved by *de-novo* assemblers (e.g. Shasta) when sequencing reads are smaller than the size of genomic fragments (mean length $>$ 125 kb in our cohort) within an ecDNA element. Decoil uses *LASSO* regression to reconstruct distinct ecDNA elements with overlapping genomic footprint, which enables the exploration of ecDNA structural heterogeneity. We have chosen this approach as it performed reasonably in our hands compared to other linear regression models (Suppl. Fig. S6). One limitation of our methods represent the correct decomposition into distinct ecDNA elements for structures containing repetitive regions. This would lead to incomplete structural resolution, e.g. the order of the repeat-containing genomic segments might remain ambiguous. Furthermore, ecDNA present at low abundance or SVs not detected

due to computational limits may affect Decoil's performance. Measuring the limit of detection of Decoil was not addressed in this manuscript, as it will require comprehensive tumor datasets with validated ecDNA structures. Ultra-long read sequencing (>100 kb) at high coverage, or other sequencing technologies, may improve the SV detection and structural resolution of ecDNA using Decoil, but aforementioned scenarios may remain difficult to resolve.

A structure-function relationship was first demonstrated for ecDNA by reports describing regulatory elements on ecDNA [5, 9, 19, 21]. These reports revealed that complex ecDNA rewire tissue-specific enhancer elements to sustain high oncogene expression [5, 22]. This also occurs through formation of new topologically associated domains [5]. Decoil was able to identify multi-region ecDNA elements, which were previously linked to enhancer hijacking [5], suggesting that it may help map such alterations in cancer. We envision that combining Decoil with DNA methylation analysis from the same nanopore sequencing reads may enable exploration of potential regulatory heterogeneity in co-occurring ecDNA elements, which was not previously possible.

The reconstruction of ecDNA in a cohort of neuroblastoma tumors and cell lines using Decoil suggested that structurally simple ecDNA elements occurred at higher copy numbers and were larger in size compared to complex ecDNA. This might be due to computational biases, as complex structures are more difficult to reconstruct, and certainly needs to be verified in larger tumor cohorts. However, it is reasonable to speculate that ecDNA complexity could influence ecDNA maintenance or impact its copy number in yet unidentified ways. Future analyzes using Decoil may help verify this observation and address such questions.

In summary, we envision that Decoil will advance the exploration of ecDNA structural heterogeneity in cancer and beyond, which is essential to better understand mechanisms of ecDNA formation and its structural evolution and may serve as the basis to identify DNA elements required for oncogene regulation and ecDNA maintenance.

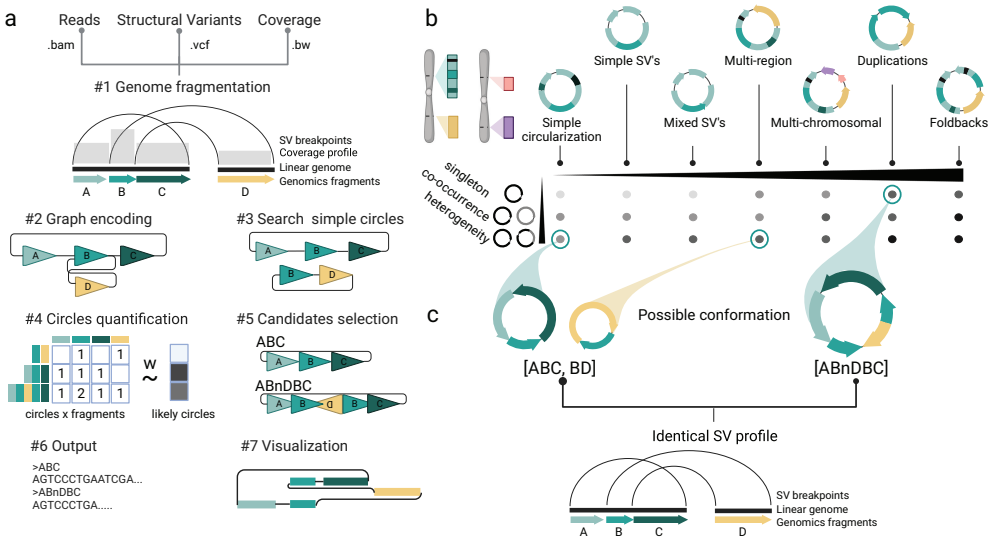

Fig. 1 Decoil algorithm overview and an ecDNA ranking system based on its structural diversity. (a) Schematic of the Decoil algorithm depicting the major steps (#1 - Genome fragmentation, #2 - Graph encoding, #3 - Search simple circles, #4 - Circles quantification, #5 - Candidates selection, #6 - Output and #7 - Visualization). (b) EcDNA diversity. The X-axis displays seven ecDNA topologies (e.g. Simple circularization, Multi-region, Multi-chromosomal) with increasing computational complexity from sequencing data. The Y-axis displays different ecDNA scenarios within one sample, i.e. singleton (presence of a single ecDNA structure), co-occurrence (presence of different ecDNA species, with non-overlapping genomic regions), heterogeneity (presence of different ecDNA species, with overlapping genomic regions). The gradient matrix summarizes the computational increasing difficulty of ecDNA reconstruction, from simple (light-gray) to very complex (black) structures captured by Decoil. (c) The overlapping cycles challenge. The left panel displays a heterogeneity scenario, where two different ecDNA elements share a genomic footprint (B fragment), the right panel displays a large structure containing interspersed-duplication rearrangement. nD - annotates inverted D fragment. Both scenarios lead to the same SV breakpoint profile. To infer the likely conformation we perform step #4, Fig. 1a. Created with BioRender.com.

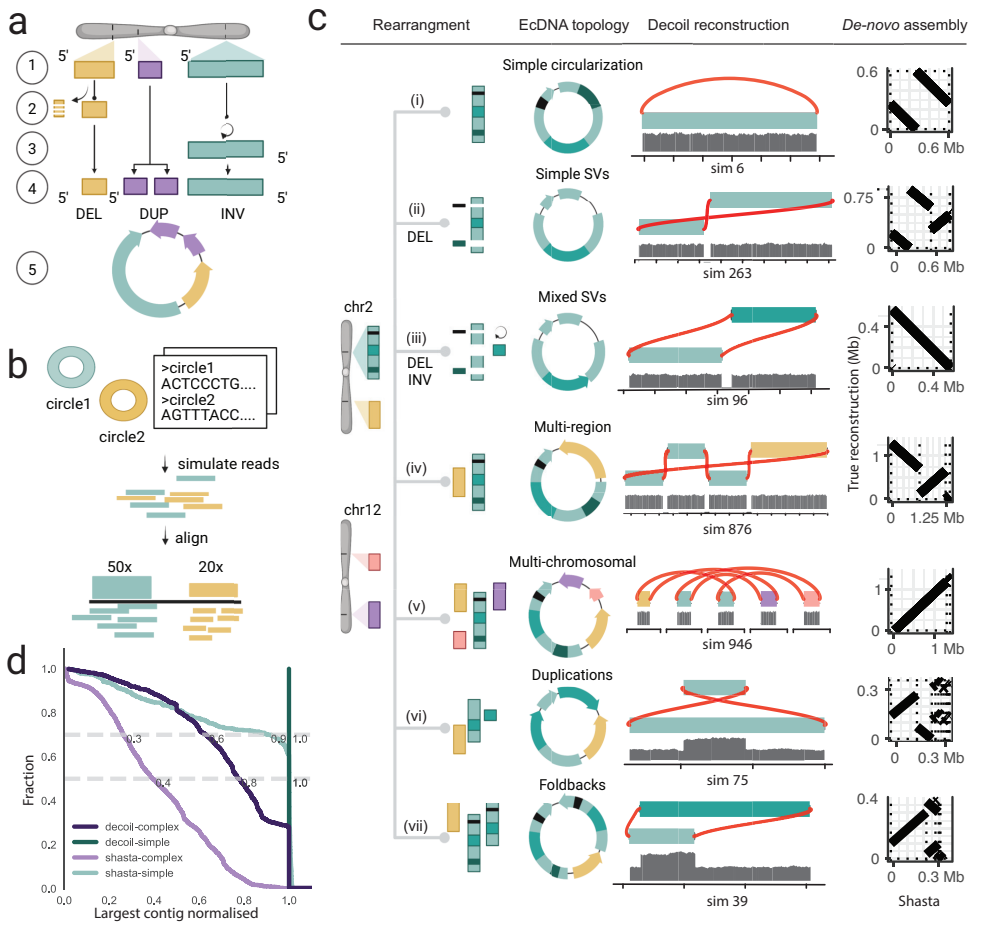

Fig. 2 Decoil reconstructs complex ecDNA elements with high fidelity from simulated data. (a) Simulation strategy for individual ecDNA templates generation, describing the main steps: 1 - choose genomic position, 2 - simulate small deletions (DEL), 3 - simulate inversion (INV), 4 - simulate tandem-duplication (DUP), 5 - generate DNA sequence template. The example shows ecDNA template containing three genomic fragments and different structural variants, i.e. 1xDEL (yellow), 1xDUP (purple), and 1xINV (green). (b) *In-silico* long-reads simulation pipeline, based on one or more ecDNA templates, at different depth of coverage. (c) EcDNA topologies, ranked with increased computational complexity, covering different simple SV mixtures: i - Simple circularization (no rearrangement on the ecDNA element), ii - Simple SV's (either DEL's or INV' series allowed), iii - Mixed SV's (mixtures of DEL's and INV's), iv - Multi-region (DEL's, INV's, TRA's mixtures originating from a single chromosome), v - Multi-chromosomal (DEL's, INV's, TRA's mixtures with fragments from multiple chromosome), vi - Duplications (DUP's + other simple rearrangements), vii - Foldbacks (INVDUP's + all other simple SV's). For each topology we show the Decoil ecDNA reconstruction together with the read coverage track. The right panel displays the *de-novo* assembly performed by Shasta (X-axis) against the true structure (Y-axis). d) Decoil and Shasta assembly contiguity for simple (i, ii, iii, iv and v topologies) and complex topologies (vi and vii). X-axis represents the larger contig normalized by the true structure length (1 - a good reconstruction, 0 - poor reconstruction, values > 1 refer to reconstructions larger than the true structure) and Y-axis shows the fraction of reconstructions with the specific contiguity. The gray horizontal lines are at 0.5 and 0.7 fraction.

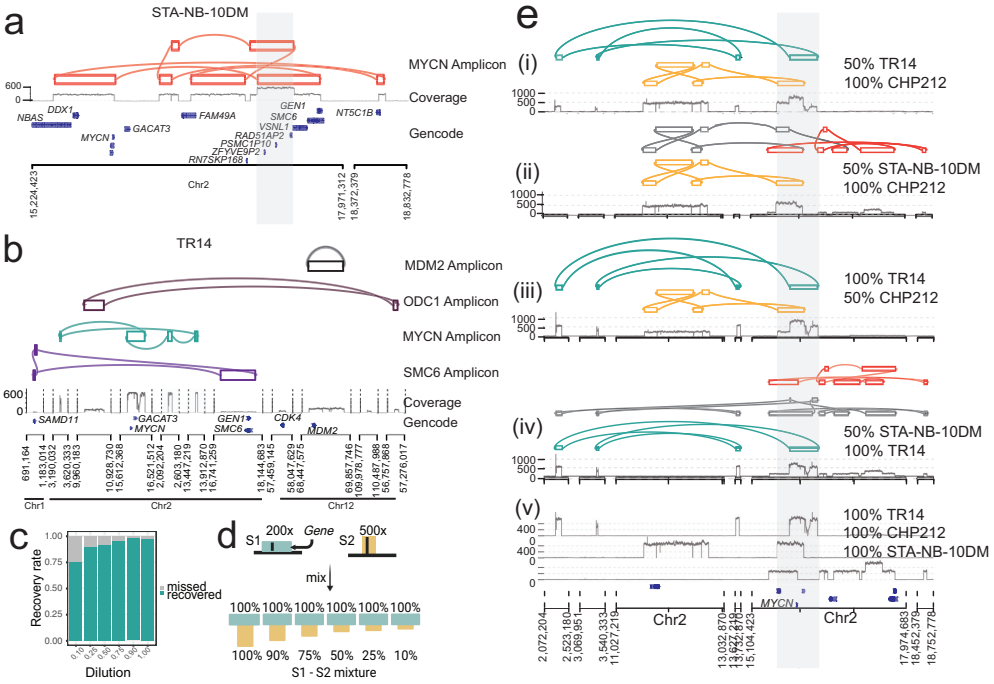

Fig. 3 Decoil captures the ecDNA structure complexity and heterogeneity in neuroblastoma cell lines. (a) STA-NB-10DM amplicon reconstruction by Decoil (top), coverage track (middle) of the aligned reads to reference genome GRCh38/hg38 and GENCODE v42 annotation (bottom). The grey highlighted region $chr2:17221081-17538185$ (GRCh38/hg38) is an interspersed duplication, covering *RAD51AP2*, *PSMC1P10* and *ZFYVE9P2* genes. (b) TR14 amplicons co-occurrence reconstructed by Decoil (top four tracks), together with the coverage track (middle) and GENCODE V42 annotation (bottom). Created with BioRender.com. (c) Recovery rate of the amplicon breakpoints (Y-axis) for *in-silico* ecDNA mixtures, in the different dilutions (X-axis). Every dot in the plot is a breakpoint which is present in the reconstruction (green) or missed (grey). The *MYCN*-amplicon for CHP212, TR14 and STA-NB-10DM is composed of 10, 8 and 14 breakpoints, respectively. For TR14 we included all ecDNA breakpoints, originating from *MYCN*, *ODC1* (4 breakpoints), *MDM2* (2 breakpoints) and *SMC6* (6 breakpoints) amplicons. (d) Dilutions strategy. Mix 100% of one sample with a fraction (10%, 25%, 50%, 75%, 90%) of another sample. Use TR14, STA-NB-10DM and CHP212 with known ecDNA structure to create the dilutions. (e) Examples of ecDNA reconstruction by Decoil for *in-silico* ecDNA mixtures. TR14 (green) and CHP212 (yellow) recovered *MYCN* amplicons in a (i) 50% to 100% and (iii) 100% to 50% mixture. (ii) STA-NB-10DM (orange) and CHP212 (yellow) recovered *MYCN* ecDNA in a 50% to 100% mixture. (iv) STA-NB-10DM (orange) and TR14 (green) recovered *MYCN* ecDNA in a 50% to 100% mixture. (v) Coverage track for pure TR14, CHP212 and STA-NB-10DM samples, at 100%. Grey amplicon regions are misassemblies. The grey shadow highlights the overlapping genomic region of the amplicons containing *MYCN*.

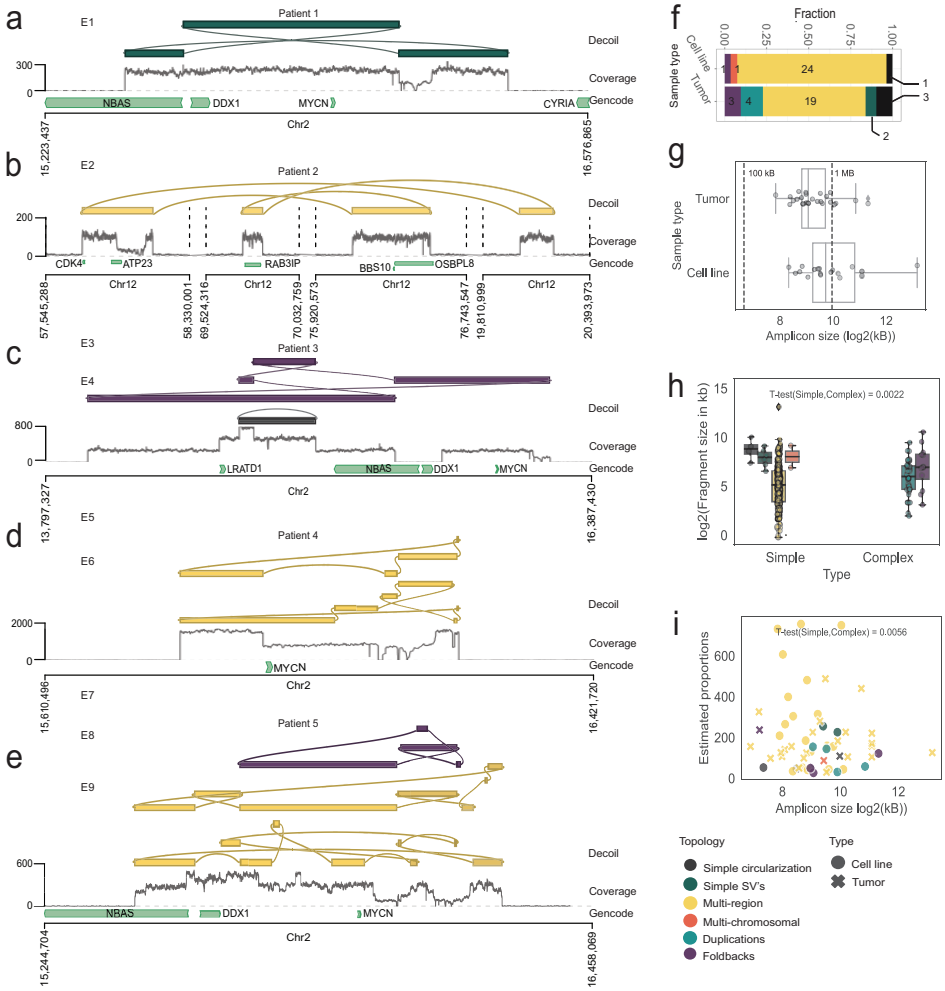

Fig. 4 Decoil recovers structurally complex ecDNA elements in primary cancers. Examples of ecDNA structure reconstruction of (a) Simple SV's, (b, d) Multi-region, (c) Foldbacks and (e) Duplication/Foldbacks topologies in patient samples. For (a-e) from top to bottom are the Decoil reconstruction, nanopore coverage of the aligned reads to reference genome GRCh38/hg38 and GENCODE v42 annotation tracks. Top 3 reconstruction per sample, labeled as ecDNA, with an estimated proportions >= 30 copies were included. E1-E9 are the ids for each reconstruction (Suppl. Table S4). (f) Decoil ecDNA reconstructions topologies fractions found in 5 cell lines and 9 patient samples. The numbers represent absolute counts. (g) Amplicon size distribution (X-axis) for cell lines and patient samples. (h) ecDNA fragment size distribution split for simple (Simple circularization, Simple SV's, Multi-region, Multi-chromosomal) and complex (Duplications, Foldbacks) topologies. T-test statistics was applied for the the fragment size of simple and complex topologies. (i) Amplicon size (X-axis) against estimated proportions (Y-axis) displayed by Decoil. T-test statistics was applied for the estimated proportions of simple and complex topologies. Amplicons labeled as ecDNA and with an estimated proportions >= 30 copies were included in (f,g,h,i). Boxplots (g,h) show Q1(25%), Q2(median) and Q3(75%), interquartile range IQR = Q3 - Q1, and whiskers are 1.5 x IQR. Colors in (a-f, h) correspond to the legend in panel (i).

4 Methods

4.1 Decoil algorithm

Decoil (deconvolve extrachromosomal circular DNA isoforms from long-read data) is a graph-based method to reconstruct circular DNA variants from shallow long-read WGS data. This uses the (1) structural variants (SV) and (2) focal amplification information to reconstruct circular ecDNA elements. The algorithm consists of six modules: genome fragmentation, graph encoding, search simple circles, circles quantification, candidates selection, output, and visualization.

Genome fragmentation

The SVs are filtered based on multiple criteria. Only SVs flagged as 'PASS' or 'STRANDBIAS', having on target coverage >= 5X (default) and VAF (Variant Allele Frequency) >= 0.2 (default) are kept. Breakpoints in a window size of 50 bp are merged. This curated breakpoints set s is used to segment the genome into $n + 1$ non-overlapping fragments $f \in F$, where F represents the non-overlapping fragments set.

Graph encoding

The coverage profile, read alignment data and fragments set F are combined to build a weighted undirected multigraph, denoted as $G = (V, E)$. In G, a vertex f represents a genomic fragment from the set F, and an edge e represents a SV connecting two fragments. Fragments with a mean coverage <= 5X (default) or standalone ($degree(v) = 0$) are discarded from the graph.

Search simple circles

Decoil continues by searching all simple circular paths c in the graph G using weighted depth-first search (DFS) approach. A cycle in a DFS tree is defined as a path where two visited nodes, u and v, are connected through a backedge (u, v), with u being the ancestor of v. This approach conducts a genome-wide search for circular paths. Duplicated cycles are removed during tree exploration. The final set S contains unique simple cycles, allowing for shared sub-paths. Simple overlapping circular paths $c \in S$ (>= 1 overlapping genomic fragment) are grouped into M non-overlapping clusters.

Circles quantification

To allow reconstruction of complex structures, e.g. containing large duplications, a set of derived cycles (D) was created. To distinguish between true possible circular DNAs and artifacts a $LASSO$ regression is fitted against targets Y, with input X, where $x_{jik} \in X$ is the occurrence of fragment f_{jk} in circle c_{ik} and $y_{jk} \in Y$ represents the total mean coverage spanning fragment j, belonging to cluster m_k. The obtained $LASSO$ coefficient represent the estimated proportions of each cycle c_{ik}. The higher the value the more likely is the cycle to be a true ecDNA element.

Candidates selection

From the candidates list we filter out cycles with an estimated proportions <= mean WGS coverage (default). Lastly, circular elements larger than 0.1 MB (threshold published by [12]) are labeled as ecDNA.

Output and Visualization

The algorithm outputs the candidates list as *.bed, *.fasta, including the mean coverage and orientation per fragment, estimated proportions of circular element. The *summary.txt* displays all found circular elements, which includes small circles and ecDNA. The reconstructions labeled as ecDNA are visualised using gGnome (https://github.com/mskilab/gGnome).

4.2 DNA extraction and nanopore sequencing

High molecular weight (HMW) DNA was extracted from 5 to 10 million cells or 15 to 25 mg of tissue using the MagAttract HMW DNA kit (Qiagen N.V., Venlo, Netherlands) according to the manufacturer's protocol. DNA concentration was measured with a Qubit 3.0 Fluorometer (Thermo Fisher) and quality control was performed using a 4200 TapeStation System (Agilent Technologies, Inc., Santa Clara, CA). For library preparation, the Ligation Sequencing Kit (SQK-LSK109 or SQK-LSK110, Oxford Nanopore Technologies Ltd, Oxford, UK) was used. All libraries were sequenced on a R9.4.1 MinION flowcell (FLO-MIN106, Oxford Nanopore Technologies Ltd, Oxford, UK) for more than 24 h.

4.3 Ranking ecDNA topologies definition

To assess Decoil's reconstruction performance, we generated an *in-silico* collection of ecDNA elements, spanning various sequence complexities for systematic evaluation. We introduced a ranking system and defined seven topologies of increasing computational complexity, based on the SV's contained on the ecDNA element: (1) *Simple circularization* - no structural variants on the ecDNA template, (2) *Simple SV's* - ecDNA contains either a series of inversions or deletions, (3) *Mixed SV's* - ecDNA has a combination of inversions and deletions, (4) *Multi-region* - ecDNA contains different genomic regions from the same chromosome (DEL, INV and TRA allowed), (5) *Multi-chromosomal*- ecDNA originates from multiple chromosomes (DEL, INV and TRA allowed), (6) *Duplications* - ecDNA contains duplications defined as a region larger than 50 bp repeated on the amplicon (DUP's + other simple rearrangements), (7) *Foldbacks* - ecDNA contains a foldback defined as a two consecutive fragments which overlap in the genomic space, with different orientations (INVDUP's + all other simple SV's). Every topology can contain a mixture of all other low-rank topologies.

4.4 Simulate ecDNA sequence templates

The simulation framework contains probabilistic variables, which model the chromosome weights, fragment position, fragment length, small deletion ratio, inversion ratio,

foldback ratio, and tandem-duplication ratio. To cover a wide range of possible conformations we generate more than 2000 ecDNA sequence templates. See Extended Methods for detailed description.

4.5 Simulate *in-silico* long-read ecDNA-containing samples

To assess ecDNA reconstruction performance, *in-silico* ecDNA-containing samples were generated based on the ecDNA sequence templates collection. The workflow takes as input the defined ecDNA elements in .bed format and generates its associated .fasta reference. Afterwards, noisy long-reads, with an average length of 7,000 bp, are sampled from this reference using an adapted version of PBSIM2 (Ono et al. 2021 [23]), at a specified depth of coverage. This package was customized for the purpose of this paper to (1) allow reads sampling from a circular reference, and (2) provide a better coverage uniformity of the reads at fragments boundary by using Mersenne twister (Harase 2014 [24]) instead of the pseudorandom number generator included in the original package (https://github.com/madagiurgiu25/pbsim2). The *in-silico* reads are stored in .fastq format. This workflow steps is part of the benchmarking pipeline https://github.com/madagiurgiu25/ecDNA-simulate-validate-pipeline.

4.6 Performance evaluation on simulated data

The *in-silico* ecDNA-containing samples were used to assess the reconstruction performance for both, Decoil and Shasta. To reconstruct the ecDNA the reads were pre-filtered using NanoFilt [25] 2.6.0 (-l 300 -q 20 –headcrop 20 –tailcrop 20). To reconstruct simulated ecDNA using *de-novo* assembly Shasta [17] 0.10.0 was used with parameters –config Nanopore-May2022 –Reads.minReadLength 1000 –Kmers.distanceThreshold 500 –Kmers.probability 0.5. To reconstruct ecDNA using Decoil the samples were preprocessed, i.e. reads were aligned to the reference genome GRCh38/hg38 using ngmlr [26] 0.2.7 with stardard parameters, structural variant calling was performed using sniffles [26] 1.0.12 (–min_homo_af 0.7 –min_het_af 0.1 –min_length 50 –cluster –min_support 4) and the bigWig coverage tracks were computed using bamCoverage (-50 bins) from deepTools [27] 3.5.1 suite. Afterwards, Decoil was applied with the parameters –min-vaf 0.01 –min-cov-alt 6 –min-cov 8 –max-explog-threshold 0.01 –fragment-min-cov 10 –fragment-min-size 500. To evaluate the correctness of reconstruction for both, Decoil and Shasta, Quast [28] 5.2.0 was applied to compute different metrics (https://quast.sourceforge.net/docs/manual.html). To overall reconstruction performance was quantified as the mean and standard deviation of the largest contig metric.

4.7 Evaluate amplicon's breakpoints recovery in ecDNA mixtures

To evaluate how well we reconstruct amplicons with overlapping footprints we generate a series of dilutions by mixing the CHP212, STA-NB-10DM and TR14 cell lines at different ratios. We generated two types of mixtures. First, we combine 100% of one sample with different percentages of another sample, i.e. 10, 25, 50, 75, 90, 100% (Figure 3c). Secondly, we generate mixtures at different ratios for both samples (10-90,

25-75, 50-50, 75-25, 90-10%). Picard 2.26 (https://broadinstitute.github.io/picard/) was used to downsample the .bam file to 10, 25, 50, 75, 90% and samtools 1.9 to merge the different ratios to create *in-silico* ecDNA mixture. SV calling was performed using sniffles [26] 1.0.12 with same parameters as for the original 100% .bam files, i.e. –min_homo_af 0.7 –min_het_af 0.1 –min_length 50 –cluster –min_support 4. Decoil was run on all these mixtures with parameters –min-vaf 0.01 –min-cov-alt 10 –min-cov 10 –max-explog-threshold 0.01 –fragment-min-cov 10 –fragment-min-size 500. The completeness of the reconstructed ecDNA elements in mixtures was evaluated by counting how many breakpoints are identical compared to the true ecDNA elements in the 100% samples.

4.8 Preprocess nanopore sequencing data from cell lines and patient samples

The cell lines CHP212, TR14, STA-NB-10DM, and all patient samples were preprocessed by performing base-calling using Guppy 5.0.14 (dna_r9.4.1_450bps_hac model), followed by a quality check using NanoPlot 1.38.1. The reads were filtered by quality using NanoFilt [25] 2.8.0 (-l 300 –headcrop 50 –tailcrop 50) and aligned using ngmlr [26] 0.2.7 against the reference genome GRCh38/hg38. The structural variant calling was performed using sniffles [26] 1.0.12 (–min_homo_af 0.7 –min_het_af 0.1 –min_length 50 –min_support 4). The bigWig coverage tracks were obtained by applying bamCoverage (-50 bins) from deepTools [27] 3.5.1 suite. The cell lines LAN-5 and CHP126 were similarly processed using the reference genome GRCh37/hg19. The pipeline is available under https://github.com/henssen-lab/nano-wgs.

4.9 Reconstruct ecDNA elements for cell lines and patient samples using Decoil

To reconstruct the ecDNA elements for CHP212, TR14 and STA-NB-10DM Decoil was applied using the parameters –min-vaf 0.1 –min-cov-alt 10 –min-cov 8 –fragment-min-cov 10 –fragment-min-size 1000 –filter-score 35 or –min-vaf 0.01 –min-cov-alt 10 –min-cov 10 –max-explog-threshold 0.01 –fragment-min-cov 10 –fragment-min-size 500, the reference genome GRCh38/hg38 and annotation GENCODE v42. Similarly, for LAN-5 and CHP126 the ecDNA reconstruction was performed using Decoil with same parameters, reference genome GRCh19/hg19 and annotation GENCODE v41. The ecDNA elements in patient samples were reconstructed by Decoil using –min-vaf 0.1 –min-cov-alt 10 –min-cov 30 –max-explog-threshold 0.01 –fragment-min-cov 20 –fragment-min-size 100.

4.10 Patient sample and clinical access

Patients were registered and treated according to the trial protocols of the German Society of Pediatric Oncology and Hematology (GPOH). This study was conducted in accordance with the World Medical Association Declaration of Helsinki (2013) and good clinical practice; informed consent was obtained from all patients or their guardians. The collection and use of patient specimens was approved by the institutional review boards of Charité-Universitätsmedizin Berlin and the Medical Faculty,

University of Cologne. Specimens and clinical data were archived and made available by Charité-Universitätsmedizin Berlin or the National Neuroblastoma Biobank and Neuroblastoma Trial Registry (University Children's Hospital Cologne) of the GPOH. The MYCN gene copy number was determined as a routine diagnostic method using FISH.

5 Acknowledgements

We would like to acknowledge Roland F. Schwarz, Julia Markowski, and Svenja Mehringer for their input and thoughtful suggestions during the development of this paper. We thank the Berlin Institute of Health (BIH) team for the support and providing the necessary infrastructure. Computation was performed on the HPC for Research cluster of the BIH. We thank the patients and their parents for granting access to the tumor specimens and clinical information that were analyzed in this study. We thank the Neuroblastoma Biobank and Neuroblastoma Trial Registry (University Children's Hospital Cologne) of the GPOH for providing samples.

6 Declarations

6.1 Funding

This project has received funding from the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme (grant no. 949172). A.G.H. is supported by the Deutsche Forschungsgemeinschaft (DFG) (grant no. 398299703). A.G.H. is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, 398299703). A.G.H. is supported by the Deutsche Krebshilfe (German Cancer Aid) Mildred Scheel Professorship program – 70114107. This project received funding from the NIH/CRUK (398299703, the eDynamic Cancer Grand Challenge).

6.2 Competing interests

A.G.H. is founder of Econic Biosciences Ltd.

6.3 Ethics approval

Patients were registered and treated according to the trial protocols of the German Society of Pediatric Oncology and Hematology (GPOH). This study was conducted in accordance with the World Medical Association Declaration of Helsinki (2013) and good clinical practice; informed consent was obtained from all patients or their guardians. The collection and use of patient specimens was approved by the institutional review boards of Charité-Universitätsmedizin Berlin and the Medical Faculty, University of Cologne. Specimens and clinical data were archived and made available by Charité-Universitätsmedizin Berlin or the National Neuroblastoma Biobank and Neuroblastoma Trial Registry (University Children's Hospital Cologne) of the GPOH. The *MYCN* gene copy number was determined as a routine diagnostic method using FISH.

6.4 Consent to participate

Not applicable

6.5 Consent for publication

Not applicable

6.6 Availability of data and materials

The sequencing data generated in this study are available at the European Genome-phenome Archive under accession no. XXXX. All other data are available from the corresponding author upon reasonable request. Source data are provided with this paper.

7 Code availability

With this article we publish several associated tools.
Decoil: https://github.com/madagiurgiu25/decoil-pre
Simulate ecDNA sequence based on specified topology:
https://github.com/madagiurgiu25/ecDNA-sim
Simulate long-reads (adapted PBSIM2 for circular reference):
https://github.com/madagiurgiu25/pbsim2
Benchmarking pipeline for *in-silico* long-read samples:
https://github.com/madagiurgiu25/ecDNA-simulate-validate-pipeline

References

[1] Kim, H. *et al.* Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nature Genetics* **52** (2020).

[2] Storlazzi, C. T. *et al.* MYC-containing double minutes in hematologic malignancies: Evidence in favor of the episome model and exclusion of MYC as the target gene. *Human Molecular Genetics* **15** (2006).

[3] Shoshani, O. *et al.* Chromothripsis drives the evolution of gene amplification in cancer. *Nature* **591** (2021).

[4] Yi, E., Chamorro González, R., Henssen, A. G. & Verhaak, R. G. Extrachromosomal DNA amplifications in cancer (2022).

[5] Helmsauer, K. *et al.* Enhancer hijacking determines intra- and extrachromosomal circular MYCN amplicon architecture in neuroblastoma (2019).

[6] Hung, K. L. *et al.* Targeted profiling of human extrachromosomal DNA by CRISPR-CATCH. *Nature Genetics* (2022).

[7] Chamorro González, R. *et al.* Parallel sequencing of extrachromosomal circular DNAs and transcriptomes in single cancer cells. *Nature Genetics* (2023).

[8] Verhaak, R. G., Bafna, V. & Mischel, P. S. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution (2019).

[9] Koche, R. P. *et al.* Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma (2020).

[10] Prada-Luengo, I., Krogh, A., Maretty, L. & Regenberg, B. Sensitive detection of circular DNAs at single-nucleotide resolution using guided realignment of partially aligned reads. *BMC Bioinformatics* **20** (2019).

[11] Zhang, P., Peng, H., Llauro, C., Bucher, E. & Mirouze, M. ecc_finder: A Robust and Accurate Tool for Detecting Extrachromosomal Circular DNA From Sequencing Data. *Frontiers in Plant Science* **12** (2021).

[12] Deshpande, V. *et al.* Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nature Communications* **10** (2019).

[13] Luebeck, J. *et al.* AmpliconReconstructor integrates NGS and optical mapping to resolve the complex structures of focal amplifications. *Nature Communications* **11** (2020).

[14] Wanchai, V. *et al.* CReSIL: accurate identification of extrachromosomal circular DNA from long-read sequences. *Briefings in Bioinformatics* **23** (2022).

[15] Olsen, N. D. *et al.* precisionFDA Truth Challenge V2: Calling variants from short- and long-reads in difficult-to-map Regions. *bioRxiv* (2020). URL https://www.cell.com/cell-genomics/pdf/S2666-979X(22)00058-1.pdf.

[16] Olson, N. D. *et al.* Variant calling and benchmarking in an era of complete human genome sequences. *Nature Reviews Genetics* (2023).

[17] Shafin, K. *et al.* Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology* **38** (2020).

[18] Storlazzi, C. T. *et al.* Gene amplification as doubleminutes or homogeneously staining regions in solid tumors: Origin and structure. *Genome Research* **20** (2010).

[19] Hung, K. L. *et al.* ecDNA hubs drive cooperative intermolecular oncogene expression. *Nature* **600**, 731–736 (2021).

[20] Pecorino, L. T., Verhaak, R. G., Henssen, A. & Mischel, P. S. Extrachromosomal DNA (ecDNA): an origin of tumor heterogeneity, genomic remodeling, and drug resistance (2022).

[21] Morton, A. R. *et al.* Functional Enhancers Shape Extrachromosomal Oncogene Amplifications. *Cell* **179** (2019).

[22] Wu, S. *et al.* Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature* **575** (2019).

[23] Ono, Y., Asai, K. & Hamada, M. PBSIM2: A simulator for long-read sequencers with a novel generative model of quality scores. *Bioinformatics* **37** (2021).

[24] Harase, S. On the F 2-linear relations of Mersenne Twister pseudorandom number generators. *Mathematics and Computers in Simulation* **100** (2014).

[25] De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics* **34** (2018).

[26] Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods* **15** (2018).

[27] Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* **44** (2016).

[28] Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. *Versatile genome assembly evaluation with QUAST-LG*, Vol. 34 (2018).