CoRAL accurately resolves extrachromosomal DNA genome structures with long-read sequencing

Kaiyuan Zhu[1,10], Matthew G. Jones[2,10], Jens Luebeck[1], Xinxin Bu[3], Hyerim Yi[2,4], King L. Hung[2], Ivy Tsz-Lo Wong[5,6], Shu Zhang[2,5,7], Paul S. Mischel[5,6], Howard Y. Chang[2,4,8,*], and Vineet Bafna[1,9,*]

[1]Department of Computer Science & Engineering, UC San Diego, La Jolla, CA, USA
[2]Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA, USA
[3]Bioinformatics Undergraduate Program, School of Biological Sciences, UC San Diego, La Jolla, CA, USA

[4]Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA
[5]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA
[6]Sarafan Chemistry, Engineering, and Medicine for Human Health (Sarafan ChEM-H), Stanford University, Stanford, CA, USA
[7]Department of Dermatology, Stanford University School of Medicine, Stanford, CA, USA
[8]Department of Genetics, Stanford University, Stanford, CA, USA
[9]Halıcıoğlu Data Science Institute, UC San Diego, La Jolla, CA, USA
[10]These authors contributed equally to this work.
[*]Correspondence: vbafna@ucsd.edu, howchang@stanford.edu

Abstract

Extrachromosomal DNA (ecDNA) is a central mechanism for focal oncogene amplification in cancer, occurring in approximately 15% of early-stage cancers and 30% of late-stage cancers. EcDNAs drive tumor formation, evolution, and drug resistance by dynamically modulating oncogene copy-number and rewiring gene-regulatory networks. Elucidating the genomic architecture of ecDNA amplifications is critical for understanding tumor pathology and developing more effective therapies.

Paired-end short-read (Illumina) sequencing and mapping have been utilized to represent ecDNA amplifications using a breakpoint graph, where the inferred architecture of ecDNA is encoded as a cycle in the graph. Traversals of breakpoint graph have been used to successfully predict ecDNA presence in cancer samples. However, short-read technologies are intrinsically limited in the identification of breakpoints, phasing together of complex rearrangements and internal duplications, and deconvolution of cell-to-cell heterogeneity of ecDNA structures. Long-

read technologies, such as from Oxford Nanopore Technologies, have the potential to improve inference as the longer reads are better at mapping structural variants and are more likely to span rearranged or duplicated regions.

Here, we propose CoRAL (Complete Reconstruction of Amplifications with Long reads), for reconstructing ecDNA architectures using long-read data. CoRAL reconstructs likely cyclic architectures using quadratic programming that simultaneously optimizes parsimony of reconstruction, explained copy number, and consistency of long-read mapping. CoRAL substantially improves reconstructions in extensive simulations and 10 datasets from previously-characterized cell-lines as compared to previous short and long-read based tools. As long-read usage becomes wide-spread, we anticipate that CoRAL will be a valuable tool for profiling the landscape and evolution of focal amplifications in tumors.

Introduction

Oncogene amplification is one of the most common events in tumorigenesis contributing to tumor initiation and progression (Beroukhim et al. 2010; Steele et al. 2022). Often, these amplifications are mediated by the formation of circular, megabase-scale extrachromosomal DNA (ecDNA) (Turner et al. 2017; Wu et al. 2019; Kim et al. 2020). Previous studies have underscored the importance of ecDNA in driving tumor formation (Luebeck et al. 2023), evolution (Lange et al. 2022), oncogene-mediated gene regulation (Hung et al. 2021; Zhu et al. 2021), and drug resistance (Nathanson et al. 2014; Lange et al. 2022). Thus, profiling the genetic and structural landscape of small, focal amplifications (typically < 10Mb), such as ecDNA, in tumors is critical for understanding the mechanisms of tumor progression and developing more effective therapies.

Owing to the large and complex genomes of ecDNA, it remains challenging to accurately infer the set of "amplicon" structures present in tumors (Deshpande et al. 2019; Luebeck et al. 2020; Chapman et al. 2023). Existing approaches rely on paired-end short-read (Illumina) sequencing to identify amplicons from copy number profiles and breakpoints that then can be represented with an edge-weighted *breakpoint graph*; ecDNAs can subsequently be extracted as cycles from the breakpoint graph (Bafna and Pevzner 1996; Alekseyev and Pevzner 2009; Lin et al. 2014; Deshpande et al. 2019; Hadi et al. 2020). Despite the success of these approaches in predicting ecDNA presence in cancer samples (Deshpande et al. 2019; Kim et al. 2020; Luebeck et al. 2023), short-read reconstructions have several limitations. First, short-read approaches struggle to handle the highly-rearranged nature of ecDNA and accurately detect breakpoints, especially in repetitive or low-complexity regions. Second, because ecDNA can contain multiple copies of large segments that are unique in the reference (e.g., Figure 1A), short-read data is limited in its ability to phase distant breakpoints correctly. Therefore, multiple collections of paths or cycles in the breakpoint graph can explain the increased copy-number equally well, masking the true structure (Figure 1C). Third, heterogeneity of ecDNA structures might result in multiple overlapping focal amplifications derived from the same genomic regions. To address these shortcomings of short-read technology, existing methods (e.g. AmpliconArchitect (AA) (Deshpande et al. 2019; Hung et al. 2022)) must use heuristics: for example, extracting cycles with the highest copy number iteratively from a breakpoint graph, until a large fraction of the aggregate copy number is explained. While these heuristic strategies return multiple small cycles (Figure 1C) that can later be recombined (Hung et al. 2022), they are still constrained by the intrinsic limitations of short-read technologies to identify structural variation and phase together distant breakpoints.

Long-reads have the potential to resolve these challenges. Recent research efforts utilized Oxford Nanopore reads to reconstruct simple ecDNA, building on off-the shelf *de novo* assembly (Helmsauer et al. 2020). However, *de novo* assembly methods often make choices based on underlying assumptions that do not hold: for example, they assume a diploid genome and that regions of high multiplicity are small enough to be spanned by long-reads. However, the heterogeneity of ecDNA structures violates the assumption of ploidy and the long segments of high multiplicity (typically 10kb-1Mb) in ecDNA are infrequently spanned by a single read, unlike the repetitive regions encountered in genome assembly, such as long interspersed nucleotide elements (LINEs) that are in the 10kb range. Concurrent with our method proposed below, a new approach, Decoil (Giurgiu et al. 2023), also aims at reconstructing ecDNA structures with long reads. However, it does not separate multiple distinct focal amplifications in one tumor sample, and uses a similar "simple cycle extraction and combining" heuristic designed for short read to reconstruct ecDNAs with high multiplicity segments. An alternative methodology utilizes optical mapping (Cao et al. 2014) (OM) to sequence large (> 200kbp) DNA fragments that span a limited number of the high multiplicity regions (Luebeck et al. 2020). While good for scaffolding, these data cannot precisely detect breakpoints, identify small structural variations, or resolve non-templated sequence, and work best in conjunction with short-read methods.

Here, we propose CoRAL (Complete Reconstruction of Amplifications with Long reads), an algorithm for reconstructing ecDNA amplicon sequence and structure from long-reads (such as those from Oxford Nanopore Technologies or PacBio). CoRAL builds a distinct breakpoint graph for each focally amplified region, and extract cycles (and walks) from the breakpoint graph representing ecDNA (and the potential focally amplified genomes). In cases where the reads are not always long enough to span the high multiplicity regions, CoRAL reconstructs likely cyclic architectures using quadratic programming that simultaneously optimizes parsimony of reconstruction, explained copy number, and consistency of long-read mapping. Through extensive benchmarks on simulated data and previously-characterized cell lines, we report that CoRAL substantially improves breakpoint detection and inferring the order of complex segments on ecDNA over long-read-based Decoil (Giurgiu et al. 2023) and the short-read-based AmpliconArchitect (Deshpande et al. 2019) methods.

Results

An overview of the CoRAL method

For better exposition of the results, we first provide a brief description of the method. A pictorial overview can be found in Figure 1D and details can be found in Methods and Supplemental Methods S2-S4. CoRAL takes mapped long-reads (in BAM format) as input and begins by identifying focally amplified *seed intervals*. The seed intervals can be provided directly, or derived from whole genome CNV calls (e.g., with third party tools like CNVkit (Talevich et al. 2016)) of mapped long reads. From the CNV calls, CoRAL selects genomic segments with minimum thresholds on copy number and aggregate size as seed intervals (Supplemental Methods S2).

CoRAL uses these seed intervals to construct a copy-number-weighted breakpoint graph separately for each amplified region. The graph construction starts with exploring all *amplified intervals* connected to the seed intervals through discordant edges given by chimeric long read mappings. Once all amplified intervals are identified for each focal amplification, a graph structure is organized by CoRAL to include the genome segments (sequence edges) from the amplified intervals, the concordant edges that join neighboring genome segments, and also the discordant edges within the amplified intervals and those connecting different amplified intervals. Once the graph structure is fixed, CoRAL recomputes the *copy number* for each edge which can best explain the long read coverage on each edge, while maintaining a balance of copy number between concordant and discordant edges incident on nodes. (see Methods and Supplemental Methods S3).

As its key step, CoRAL reconstructs potential ecDNA structures in the breakpoint graph by extracting a minimum number of *cycles and walks* from the graph, allowing duplication of nodes (e.g. Figure 1C), where cycles represent the potential ecDNA species and walks represent linearly amplified or rearranged genome. Each cycle/walk is associated with a positive weight – corresponding to the copy number – so that the sum of length-weighted edges of extracted walks explains a large fraction of the total copy number of the edges in the breakpoint graph. In addition, CoRAL takes advantage of the fact that long-reads may span several breakpoints and incorporates these reads as *subwalk constraints*. In its cycle extraction, CoRAL also requires a majority of the subwalk constraints to be satisfied by the resulting cycles and walks, thus leveraging the power of long reads. CoRAL uses quadratic integer programming to solve a multi-objective optimization that minimizes the number of cycles/walks while maximizing the explained length-weighted copy number and the number of subwalk constraints (Methods and Supplemental Methods S4). It finally outputs the reconstructed breakpoint graphs for each focal amplification in the sample, as well as the associated cycles/walks from the graph. It also optionally outputs stylistic visualizations of the breakpoint graphs and cycles, as shown in subsequent results.

Simulation benchmarks

We first assessed the effectiveness of amplicon reconstruction algorithms using simulated sequencing data from synthetic amplicon structures (Supplemental Methods S5, Supplemental Table S1, S2). To capture the diversity of ecDNA amplicons observed in patient tumors and cell lines, we simulated 75 distinct cyclic structures with varying numbers of breakpoints (between 1 and 20) from one of three origins: *episomal*, in which a contiguous region of the genome is excised from a chromosome; *chromothripsis*, in which a mitotic defect leads to the shattering of a lagging chromosome and ecDNA formation (Ly et al. 2017; Shoshani et al. 2021); or, finally, *2-foldback*, in which extruding double-stranded DNA from a stalled replication fork is broken off as ecDNA (Passananti et al. 1987). Our simulated ecDNAs additionally included internal structural variants in the form of insertions, deletions, duplications and inversions (see Supplemental Methods S5 for more detailed description of the simulation process and Supplemental Table S1 for the data). Subsequently, each test dataset was generated by randomly selecting between 1 and 5 amplicon structures (from the pool of 75 synthetic amplicons). Reads from long-read (using Nanosim (Yang et al. 2017)) and Illumina short-read, paired-end technologies (using Mason (Holtgrewe 2010)) were simulated from these amplicons at one of three coverages (50×, 100×, or 250× coverage; or approximate copy-numbers of

7, 15, or 37, respectively) and merged with reads from one of five simulated normal, diploid genomes (each with ~13× coverage). A total of 50 test datasets were simulated in this fashion and used for benchmarking amplicon reconstruction (Supplemental Table S2).

From these inputs, ecDNA was reconstructed using simulated long-reads provided to CoRAL and Decoil (Giurgiu et al. 2023) - a separate long-read amplicon reconstruction tool - or simulated short-reads provided to AmpliconArchitect (AA). In most cases, the *heaviest* CoRAL cycle, defined as the cycle with the largest length-weighted copy number, was better at recapitulating the true architecture compared to the AA cycle (e.g., Figure 2A). We systematically evaluated the accuracy of the best reconstruction W_r (as defined as the highest-scoring reconstruction with respect to a particular statistic) against a true cycle W_t using four additional measures defined briefly below (Figure 2B-E; see Supplemental Methods S7 for more detailed definitions):

1. **Breakpoint Graph Accuracy** reports the proportion of discordant edges that agree, up to a tolerance of 100bp, between the true breakpoint graph G_t and reconstructed breakpoint graph G_r.
2. **Cycle Interval Overlap** measures the Jaccard index, weighted by the number of nucleotides, of the genomic intervals defined by W_t and W_r.
3. **Cyclic Longest Common Subsequence (LCS)** measures the length of the longest common subsequence contained in W_t and W_r after eliminating intervals that are not found in both, normalized to the length of W_t.
4. **Reconstruction length error** reports the difference in amplicon lengths between W_r and W_t, normalized by the true amplicon length W_t. We report log_2-scaled values.

Across the 50 simulated datasets, we observed consistently improved reconstruction of CoRAL over AA and Decoil for all four measures (Figure 2B-E). Notably, 93% of CoRAL reconstructions perfectly recapitulated the ground truth breakpoint graph as compared to 51% for Decoil and 4%

for AA (Figure 2B). These results underscore the improved mapping of structural variants with long-reads.

While CoRAL outperformed Decoil and AA in all four measures, both AA and Decoil capture many critical aspects of the amplicon, such as including the most amplified intervals (Figure 2C) and capturing the true ordering of the segments (Figure 2D). Mostly, CoRAL's improved performance is reflected in reconstructed cycle lengths that are most similar to the true cycle (Figure 2E). In addition, we observe that both AA and CoRAL tend to produce a main cycle that account for a large fraction of length-weighted copy-number (Supplemental Fig S1) and that this weight ratio is correlated with cycle reconstruction accuracy (Supplemental Fig S2). Through these analysis, we also noted several examples where the interval ordering is incorrect despite near-perfect recovery of breakpoint graph and interval overlap (Supplemental Fig S3), reflecting the technological limitations of reads that were not long enough to resolve the true order of segments.

We also compared reconstruction performance as a function of the complexity of amplicons (number of segments, or sequence edges), sequence coverage, their formation context, and level of duplication (or multiplicity). We observed that the number of segments in the true amplicon had modest effects on reconstruction accuracy, (Supplemental Fig S4), as did coverage (and by extension copy-number; Supplemental Fig S5). These observations suggest that all algorithms, but especially CoRAL, can accurately reconstruct complex ecDNAs at low copy numbers (e.g. < 7). We additionally observed that increasing levels of segmental or breakpoint multiplicity often resulted in poorer reconstruction accuracy for all methods tested, though CoRAL remained mostly robust (Supplemental Fig S6). However, in considering the various contexts in which ecDNA can form (Bafna and Mischel 2022), we observed substantial performance differences: generally, we observed that chromothripsis amplicons were most difficult for AA and CoRAL with Decoil modestly outperforming CoRAL conversely 2-foldback amplicons were most difficult for Decoil. These observations highlight the importance of accurately detecting structural variants, which is greatly enhanced with long reads, but can be nevertheless challenging depending on the complexity of breakpoints (Supplemental Fig S7).

Amplicon reconstruction in cell lines.

Next, we evaluated amplicon reconstruction using matched Nanopore long-read sequencing and Illumina paired-end short-read sequencing in 7 previously characterized cell-lines spanning a range of cancer types and amplifications (summarized in Table 1 and Supplemental Table S3): COLO320(-DM, -HSR), PC3(-DM, -HSR), GBM39(-HSR), and CHP-212. Of these 7 cell lines, there are 3 isogenic pairs in which the amplified oncogene is located on chromosomal homogeneously staining regions (HSRs) while maintaining the core cyclic structure, as opposed to ecDNA (e.g., COLO320-

HSR vs COLO320-DM). Additionally, we assessed reconstruction in 4 recently monoclonalized versions of these cell lines (PC3-DM, PC3-HSR, GBM39ec, and COLO320-DM). Together, this resulted in a matched Nanopore and Illumina data for 10 samples for analysis.

CoRAL accurately predicts the existence of ecDNA. We ran the AmpliconClassifier (Luebeck et al. 2023) method to reconfirm the cyclic structure of the ecDNA amplicons in all samples. AmpliconClassifier parses the breakpoint graph and identifies sub-graphs as being cyclic (or ecDNA), breakage fusion bridge, heavily-rearranged, or linear-rearranged (Kim et al. 2020). CoRAL identified altogether 60 amplicons in the 10 cell lines, including the main ecDNA (or HSR) amplicon in each sample. AmpliconClassifier consistently classified the main ecDNA amplicon as cyclic with the breakpoint graphs constructed by CoRAL using long-reads and AA using short-reads, indicating the existence of ecDNA (or HSR). Long-read sequencing did not identify new ecDNA amplicons nor fail to detect previously confirmed ecDNA amplicons in the cell line samples.

CoRAL cycles better explain the copy numbers in ecDNA amplicons. To benchmark the reconstruction quality of CoRAL and AA in cell lines, we computed the fraction of length weighted copy numbers in the breakpoint graph given by the *k*-heaviest cycles, which we previously observed correlated with accuracy (with k = 1), for k = 3 and k = 1, in each of the 10 ecDNA cell lines (Figure 3A, Supplemental Fig S8A; see Supplemental Methods S7 for details of the statistic). Consistent with simulated data, these results demonstrated that CoRAL explains a higher fraction of the length-weighted copy number with fewer cycles. Across all samples, the copy number explained by the 3 heaviest cycles was substantially higher for CoRAL compared to AA (Figure 3A). The reconstructed cycles of COLO320-DM (Wu2019), the monoclonal COLO320-DM (mono) and the shallow coverage COLO320-DM (Hung2021) showed consistent heaviest cycle weight ratio (Figure 3A, Supplemental Fig S 8A), shared many structural features, and contained a similar subset of genes (Figure 3C,D, Supplemental Table S4), even as they showed some differences in the reconstructed amplicons (Supplemental Fig S9). These differences could reflect differences in intrinsic heterogeneity or evolution of the cell line over time, which also resulted in lower heaviest cycle weight ratio in COLO320-DM cells and its isogeneic pair COLO320-HSR, in comparison to the GBM39 and CHP-212 cell lines with a single dominating ecDNA structure (Wu et al. 2019; Helmsauer et al. 2020) (Supplemental Fig S 8A).

CoRAL cycles satisfy more subwalk constraints. In its optimization, CoRAL takes advantage of the fact that long-reads may span several breakpoints and incorporates these reads as *subwalk contraints* that can be satisfied during cycle decomposition and lend support for accurate reconstruction. As such we mapped each long read subwalk constraint to each AA cycle and checked if the subwalk constraint can also be satisfied by that cycle. Expectedly, CoRAL satisfied more subwalk constraints compared to AA (Figure 3B, Supplemental Fig S 8B), especially for the complex amplicons. For example, CoRAL satisfied 1.5× and 25× more subwalk constraints in COLO320-DM (mono) and PC3-DM (mono), respectively. Together, these subwalk constraints support most junctions of the amplicon (e.g., see Figure 3C,D), thereby taking advantage of the long-reads that span multiple breakpoints. Nevertheless, no reconstruction satisfied all subwalk constraints in either CoRAL or AA, consistent with the high heterogeneity of ecDNA structure in samples.

CoRAL cycles enable the study of critical aspects of amplicon structures. Reconstruction supported by long-read subwalk constraints additionally enabled the study of critical aspects of the amplicon structures. As one example, Figure 3E and F show the reconstruction of the two heaviest CoRAL and the three heaviest AA cycles, respectively, for monoclonal COLO320-DM. To note, the monoclonal COLO320-DM is a recently derived line from a parental line where previous experiments integrating WGS, optical mapping, and in-vitro ecDNA digestion revealed an ecDNA structure of approximately 4.3Mb (Hung et al. 2021). Here, the automated reconstruction of monoclonal COLO320-DM using CoRAL also revealed an ecDNA of size 4.4Mbp (Figure 3D) which shared many structural features with the previous reconstruction.

One distinct feature of the COLO320-DM *MYC* amplicon is the overexpression of a fusion transcript consisting of a truncated, 5' portion of the lncRNA *PVT1* fused to the second exon of the *MYC* oncogene (Hung et al. 2021). This is despite *PVT1* being positioned downstream of *MYC* in the reference genome. As expected, CoRAL reconstruction of COLO320-DM includes a breakpoint that connects a truncated, 5' portion of *PVT1* upstream of exon 2 of *MYC*, thereby explaining the fused transcript. Notably, both CoRAL and AA detected the *PVT1* -*MYC* fusion breakpoint in all COLO320-DM samples; however, CoRAL's cycle decomposition included this breakpoint in the heaviest (largest length-weighted CN) cycle across multiple COLO320-DM samples (Figure 3C,D,E, Supplemental Table S4). AA did not include the breakpoint in the 3 heaviest cycles (Figure 3F); instead, it reports a smaller cycle of size ~90 kbp containing the breakpoint by itself. Furthermore, subwalk constraints due to long-reads linked truncated *PVT1* and *MYC* on a single molecule. Correspondingly, CoRAL reconstructions of cycles in COLO320-DM (mono), COLO320-DM (Hung 2021), and COLO320-DM (Wu 2019) all showed the three elements in a single cycle (Figure 3C,D, Supplemental Table S4).

We additionally observed that subwalk constraints and CoRAL's cycle reconstructions support a co-amplification of the ncRNA *PCAT1* and *MYC* on COLO320-DM ecDNA (Figure 3E,F). Previous

DNA FISH experiments also confirmed the co-existence of these genes on COLO320-DM (Hung 2021 (Hung et al. 2021), Extended Figure 4g). The *PCAT1* ncRNA is known to repress *BRCA2* (Prensner et al. 2014b), activate *MYC* (Prensner et al. 2014a), and promote cell proliferation (Xiong et al. 2019), and is upregulated in prostate, colorectal, and other cancers (Xiong et al. 2019). Thus, these CoRAL cycle reconstructions are also consistent with the regulatory and pro-oncogenic roles of *MYC* and *PCAT1*. Together, these results highlight the advantages of CoRAL in reconstructing complex ecDNAs in cell lines and may enable new biological insights into the co-amplification of genetic elements on the same ecDNA molecule.

CoRAL requires comparable computational resources to AA. We finally compared the computational resources required by CoRAL and AA to reconstruct all amplicons in these cell lines. To perform a fair test, we ran CoRAL and AA on the same Ubuntu system (2× Intel Xeon X5680 CPUs, and 128G RAM). Importantly, we observed that total running time and memory of CoRAL was comparable to AA for reconstructing the amplicons, even if an MIQCP was solved for each amplicon (Supplemental Fig S10). The most complex sample, COLO320-HSR (Wu 2019), was completed in less than 22h (8×10^5s) for CoRAL. Furthermore, we found that most focal amplifications except ecDNA are relatively easy to resolve, with the resulting breakpoint graphs being small – out of the 60 amplicons detected by CoRAL across all samples, only 8 required greedy MIQCP, including 7 of the 10 total ecDNA amplicons.

Discussion

Our results suggest that long-read guided reconstruction greatly improves ecDNA structure resolution, both in individual detection of breakpoints and in the accuracy of the large-scale predicted structure. The constrained optimization performed by CoRAL reconstructs plausible structures based on selecting a minimum number of cycles that are consistent with the constraints provided by long-reads, and together, the cycles explain most of the copy number of the amplicon. On simulated data, most structures were correctly predicted, and even when they were not, they were only slightly rearranged from the true structure. Similarly, in experimental data from cancer cell-lines, the three heaviest reconstructed structures typically explained most of the copy number. In most cases, the reconstruction provides a reasonable template for downstream functional studies, including analysis of regulatory rewiring and chromatin conformation. Of note, CoRAL's approach can be seamlessly employed for any long-read sequencing technology, such as Oxford Nanopore Technologies or PacBio, where longer reads will always improve breakpoint detection and amplicon reconstruction.

It is important to note, however, that long-reads by themselves are not a panacea, and amplicon reconstruction is different than genome assembly. In diploid genomes, only two haploid structures are possible, and the repeated regions are easily spanned by current long-read technology, except in a few highly repetitive regions. In contrast, larger regions can occur with multiple copies on a single ecDNA, making it hard to resolve into one correct structure. Moreover, heterogeneity of ecDNA may lead to many structures being present. The ecDNA structures resolved by CoRAL may only reflect the most abundant structures. Moreover, due to the minimization of cycle counts, it is possible that the heaviest cycle given by CoRAL glues together smaller ecDNA cycles that share the same segments. To avoid such cases we limit the times that each discordant edge can be traversed in a cycle or walk based on empirical observations. Reconstructions from simulated and real ecDNA amplicons (e.g. COLO320DM) suggested similar cycle sizes to either ground truth or previous characterizations. These considerations will be revised as additional data becomes available.

Thus, we highlight a few avenues for extending and improving CoRAL. First, when a sample has concurrent short-read sequencing data, one may explore if incorporating low-coverage long-reads ($< 5\times$) are sufficient for a hybrid reconstruction. However, due to the rapid evolution of cancer genomes and spatial heterogeneity of tumor samples, the benefit of such an approach may only exist when short and long reads are simultaneously generated from the same biospecimen. Second, CoRAL can be extended to identify the architectures of chromosomal amplicons such as breakage fusion bridge cycles, and ecDNA that have reintegrated into the genome. Because the reconstruction methods use only abstractions relating to path constraints and explained copy number, they can be adapted to other amplifications readily, and this will also be a focus of future studies. Third, as our understanding of amplicon structure grows with experimentally verified structures, that information can be used to improve the constraint space and optimization criteria for CoRAL, and to enhance the simulations of ecDNA or other chromosomal amplifications.

Previous state-of-the-art tools using short-reads like AA (Deshpande et al. 2019) are very accurate in determining if a focal amplification is mediated by ecDNA formation, and in determining the amplified regions. However, they have difficulties in reconstructing the full structure, or determining all regions that participate in one ecDNA molecule. These challenges are partially resolved by targeted deep profiling of a specific subset of amplicons at the expense of not observing the full amplification landscape (Hung et al. 2022). CoRAL not only offers improvements as a standalone tool, but can also be used in conjunction with the targeted approaches - either by refining existing reconstructions or by providing more accurate and unambiguous reconstructions of complex amplicons in targeted enrichment protocols. In summary, CoRAL will be a valuable tool in the arsenal for analyzing complex focal amplifications - such as ecDNA - in tumor genomes, especially as long-read technologies continue to offer cheaper, longer, and more accurate reads.

Methods

CoRAL takes mapped long-reads (in BAM format) as input, constructs a copy-number-weighted breakpoint graph, decomposes the breakpoint graph into a collection of cycles or paths, and outputs the reconstructed breakpoint graph as well as the resulting cycles/paths from decomposition of the breakpoint graph. A pictorial overview of CoRAL procedure is given in Figure 1D.

Below, we start with an abstract definition of the breakpoint graph followed by a high-level description of the construction. The copy-number-weighted breakpoint graph (Bafna and Pevzner 1996; Alekseyev and Pevzner 2009; Lin et al. 2014; Deshpande et al. 2019; Hadi et al. 2020), denoted by $\mathcal{G} = (V, E = E_s \cup E_c \cup E_d, \text{CN})$, encodes a collection of non-overlapping intervals on a given reference genome, which are amplified, reordered or reoriented. A brief description is provided here with details in Supplemental Methods S1:

- Each $v \in V$ represents the start or end coordinate of an interval, or the special *source* nodes s, t (defined below). Let l_v denote the location of node v.
- E_s represents *sequence edges*, that join the start and end coordinates of an interval.
- E_c represents *concordant edges* so that $(u, v) \in E_c$ if $l_v - l_u = 1$ where v is the start coordinate of the canonically larger interval on the reference genome represented by a sequence edge.
- E_d represents *discordant edges*, generated when (sufficient) reads map to discordant intervals. Thus, $(u, v) \in E_d$ if $|l_v - l_u| \neq 1$, if the read connecting u to v changes orientation, or if the nodes are on different chromosomes. A discordant edge could connect the start (or end) coordinate of an interval to itself (an inverted duplication or *foldback*).
- All edges are weighted using the real-valued function CN: $E \rightarrow \mathbb{Q}^+$ denoting the *copy number*. The CN is computed based on an assumption of diploidy for the majority of base pairs on the genome. We require that the CN assignment be "balanced", for each $(u, v) \in E_s$, as follows:

$$\sum_{(w,u) \in E_c \cup E_d} \text{CN}(w, u) = \text{CN}(u, v) = \sum_{(v,w) \in E_c \cup E_d} \text{CN}(v, w), \forall (u, v) \in E_s \qquad (1)$$

By definition, each node in a breakpoint graph is connected to a single sequence edge, and a single concordant edge as well; but it may connect to multiple discordant edges. The source nodes s and t connect to the canonically smallest and canonically largest coordinate on the reference genome from a collection of consecutive intervals connected by concordant edges; or a sequence edge which is only connected to another sequence edge with smaller CN by concordant edges, and therefore is deemed to violate the balanced CN constraint without the source connections. Edges connected to source nodes are treated as discordant edges. See Figure 1B for an example of a breakpoint graph constructed by CoRAL from the ecDNA in Figure 1A. We denote a maximal collection of genomic intervals connected by concordant edges as an *amplified interval*, and the union of all amplified intervals and their (discordant) connections as an *amplicon*. Note that a tumor sample could contain multiple amplicons whose intervals are non-intersecting. CoRAL constructs a distinct breakpoint graph for each amplicon.

Breakpoint graph construction with CoRAL. To build the breakpoint graph for an amplicon, CoRAL first determines all amplified intervals included in the amplicon. CoRAL requires *seed amplified intervals* (Supplemental Fig S 11A) as a starting point to search for all connected amplified intervals contained in an amplicon. The seed amplified intervals can be derived from whole genome CNV calls (e.g., with third party tools like CNVkit (Talevich et al. 2016)) of mapped long reads. From the CNV calls, we select the genomic segments adjacent to each other with a minimum threshold of copy number as well as the aggregated size as seed intervals (Supplemental Methods S2).

With the seed amplified intervals, CoRAL searches for amplified intervals connected to seed intervals (by discordant edges) using a breadth first search (BFS). For BFS CoRAL maintains a list $\mathcal{I}$ of amplified intervals in all amplicons it explored or discovered so far, initialized as the list of seed intervals; and a set $\mathcal{E}$ representing the connections between amplified intervals through discordant edges. Each pair of intervals $(a_i, a_j) \in \mathcal{E}$ $(i \leq j)$ is labeled by the breakpoints connecting two loci within the intervals a_i and a_j respectively. The main iteration explores the next unvisited interval a_i in $\mathcal{I}$, indicating a new amplicon (connected component), until all amplified intervals in $\mathcal{I}$ are visited. Let L be a priority queue used in the interval search starting from a_i, which is initialized with a single element a_i. Each step of the interval search pops the first interval a_0 in L, and extracts all breakpoints supported by chimeric alignments connecting a locus within a_0 to another locus on the reference genome. These breakpoints are greedily clustered (with the procedure described below) and the new locus l determined by a cluster $\mathrm{bp}_{a_0,l}$ of breakpoints of size at least haploid coverage is chosen to be further explored (Supplemental Fig S11B). If the new locus falls into an existing interval $a_e \in \mathcal{I}$, then mark interval a_e as visited, augment the label set of (a_0, a_e) with $\mathrm{bp}_{a_0,l}$, and only append a_e to L if it was not previously visited. Otherwise CoRAL will extend l to a new amplified interval including l, depending on whether l is amplified from the CNV calls. If l is amplified, CoRAL will append the new interval $a_n = [\mathrm{chr}_l, \max(s_l - \delta, l - \Delta), \min(e_l + \delta, l + \Delta)]$ to both L and $\mathcal{I}$, where s_l and e_l are the start and end coordinate of the amplified CN segments including l in CNV calls. If l is not amplified, CoRAL will append the new interval $a_n = [\mathrm{chr}_l, l - \delta, l + \delta]$ to L and $\mathcal{I}$. In either case, CoRAL also labels the connection (a_0, a_n) with $\{\mathrm{bp}_{a_0,l}\}$ and add it to $\mathcal{E}$. The amplified interval search starting from a_i is repeated until L becomes empty. A pseudocode of the above procedure as well as the selection of Δ and δ is discussed in detail in Supplemental Methods S2.

At the end of interval search, all intervals $\mathcal{I}$ are visited, and each connected component of amplified intervals by breakpoint edges with sufficient support of long reads forms an amplicon (Supplemental Fig S11C). After BFS, CoRAL postprocesses the amplified intervals discovered in $\mathcal{I}$ by merging (i) adjacent (in CNV calls) or overlapping intervals, or (ii) intervals on the same chromosome which are not adjacent but have close (i.e., within $\leq 2\delta$-bp vicinity) breakpoint connections. Two intervals belonging to different amplicons are brought into the same amplicon after merging. CoRAL will then search for breakpoints within a single (merged) amplified interval (Supplemental Fig S11D). Finally, CoRAL builds the actual breakpoint graph for each amplicon. It will split each amplification interval into sequence edges if there are breakpoint edges connecting to the middle of that interval, and add concordant edges connecting two adjacent sequence edges on the reference genome (Supplemental Fig S11E).

CN assignment. Once the graph structure $\mathcal{G}$ is fixed, CoRAL recomputes the CN for each edge in $\mathcal{G}$ (Supplemental Fig S11E), as the initial CNV calls used for amplified interval search may not follow the balance requirement (Eqn. 1), and they do not account for concordant and discordant edges. Let the diploid long read coverage be θ. CoRAL assumes that the majority of the donor genome is not amplified and estimates θ as the coverage on the 40-th percentile of CN segments sorted by their initial CNV calls. Given θ, CoRAL models the total number of nucleotides on each sequence edge $(u, v) \in E_s$ as a normal distribution with mean and variance $\theta \cdot \mathrm{CN}(u, v) \cdot l(u, v)$, where $l(u, v)$ denotes the length (in bp) of the sequence edge; and the number of reads supporting each concordant and discordant edge $(u, v) \in E_c \cup E_d$ as a Poisson with mean $\theta \cdot \mathrm{CN}(u, v)$. To estimate CN, CoRAL computes the maximum likelihood of CN using the joint distribution of observed number of nucleotides on each sequence edge and the observed read counts on each concordant/discordant edge–with the constraint that CN is balanced (Supplemental Methods S3). The (convex) optimization problem was solved using CVXOPT package.

Cycle extraction. We are interested in paths and cycles that alternate between sequence and breakpoint (i.e., concordant or discordant) edges, thus by definition, if the path contains node s (respectively, t), it must be the first (respectively, last) node in the path. Define an *alternating sequence* of nodes as a sequence $v_1, v_2, \cdots, v_w$, where for all, $1 \leq i < w$, $(v_i, v_{i+1}) \in E$ and the edges alternate between sequence and breakpoint edges. Define a *walk* in $\mathcal{G}$ as an alternating sequence $v_1, v_2, \cdots, v_w$, where $v_1 = s, v_w = t$. A *path* is a walk with no node repeated ($v_i = v_j \Leftrightarrow i = j$). A *cyclic walk* or *cycle* is an alternating sequence $v_1, v_2, \cdots, v_w$ of nodes where $v_1 = v_w \neq s, t$. The cycle is simple if no node except the first/last one is repeated.

The amplicon encoded by $\mathcal{G}$ is composed of a superposition of cycles and walks with high copy numbers. For all sequence edges $(u, v) \in E_s$, define the *length-weighted copy number* using $\mathrm{C}_l(u, v) = \mathrm{CN}(u, v) \cdot l(u, v)$. Similarly, for graph $\mathcal{G}$,

$$\mathrm{C}_l(\mathcal{G}) = \sum_{(u,v)\in E_s} \mathrm{C}_l(u, v). \qquad (2)$$

Our goal is to identify a minimum number of cycles and walks (denoted as W_i), each associated with a positive weight – corresponding to the copy number (based on the assumption of uniform coverage, see Supplemental Fig S12) – so that the sum of weights on all edges in all walks composes a large fraction of $\mathrm{C}_l(\mathcal{G})$. Furthermore, the long-reads that span multiple (at least 2) breakpoints in $\mathcal{G}$, also provide us with a collection of subwalks $\mathcal{P} = \{p_1, \cdots, p_m\}$, and the reconstructed walks must simultaneously be consistent with a large fraction of these subwalks (Supplemental Methods S4).

The complexity of cycle extraction and rationale for CoRAL's optimization procedure can be motivated through an example illustration (Supplemental Fig S13). The breakpoint graph in Supplemental Fig S13A consists of segments A, B, and C assumed for simplicity to be of equal length. The optimization in CoRAL will decompose it into a single cycle of copy number 50, with a duplication of segment B (right panel). The decomposition is also supported by the subwalk constraint given by the long read that connects segment A, B and C. Even if the long-read were not present, this cycle is still a parsimonious solution compared to an alternative decomposition with two cycles (one containing A and B, and the other containing B and C).

Supplemental Fig S13B has a similar graph structure but with different copy numbers on segments. The best decomposition is given by one cycle containing A and B with copy number 80, and a second cycle containing A, B (2 copies), C with copy number 10. The total length weighted copy number of the graph is 200 (assuming segments of length 1). Cycle 1 explains 80% and Cycle 2 explains 20% of the copy number. Other decompositions of cycles are indeed possible. For example, if the subwalk constraint given by the long read were not present, an alternative decomposition with 90 copies of Cycle 1 and 10 copies of a different cycle 2 containing only segments B and C would also be explain all length-weighted copy numbers in the graph. On the other hand, a more parsimonious decomposition into one single cycle with copy number 10, where segment A repeating 9 times and segment B repeating 10 times is not allowed because it violates the upper bound on the multiplicity of segments in the cycle (see auxiliary constraint 3 in the MIQCP formulation below). For the same reason, the decomposition into one single cycle is not allowed in Supplemental Fig S13C.

We resolve the multi-objective challenge using mixed integer quadratically constrained programing (MIQCP). The MIQCP works with 2 parameters: α as the minimum fraction of length-weighted copy number explained, and β as the minimum fraction of path constraints satisfied. Additionally, parameter k denoting the maximum number of cycles/walks allowed is learned starting with $k = 10$, according to two modes. In the **full** mode (MIQCP-full) described below, the MIQCP attempts a solution with at most k walks that satisfy other constraints, or returns 'infeasible'. The value of k is doubled until feasibility is reached or $k > |E|$. The **greedy mode** is described later. We implement both quadratic programs with through the python3 interface of Gurobi 10.0.1.

MIQCP-full utilizes the following **key** variables.

- $w_i \in \mathbb{Q} > 0$ denotes the copy number for walk W_i ($1 \leq i \leq k$); and $z_i \in \{0,1\}$ indicates if $w_i > 0$;
- $x_{uvi} \in \mathbb{Z} \geq 0$ represents the number of times walk W_i traverses (u,v) for each edge $(u,v) \in E$ and $1 \leq i \leq k$;
- $P_j \in \{0,1\}$ indicates if subwalk constraint p_j is satisfied for $1 \leq j \leq m$;

The MIQCP(k, α, β) objective is given by:

$$\min \underbrace{\sum_{i=1}^{k} z_i}_{\#\text{walks}} - \underbrace{\frac{1}{\mathrm{C}_l(\mathcal{G})} \sum_{i=1}^{k} \sum_{(u,v)\in W_i \cap E_s} w_i \cdot x_{uvi} \cdot l(u,v)}_{\text{fraction of } \mathrm{C}_l(\mathcal{G}) \text{ explained}} - \underbrace{\frac{1}{m} \sum_{j=1}^{m} P_j}_{\text{fraction of subwalks satisfied}}, \qquad (3)$$

subjects to the constraints:

$$w_i \leq z_i \cdot \mathrm{C}_l(\mathcal{G}) \qquad (4)$$

$$\sum_{i=1}^{k} \sum_{(u,v)\in W_i \cap E_s} w_i \cdot x_{uvi} \cdot l(u,v) \geq \alpha \cdot \mathrm{C}_l(\mathcal{G}) \qquad (5)$$

$$\sum_{j=1}^{m} P_j \geq \beta \cdot m \qquad (6)$$

Constraint 4 ensures that $w_i = 0$ if $z_i = 0, \forall i = 1, \cdots, k$, and constraints 5, 6 ensure that minimum fractions of the length weighted copy number and subwalk constraints are satisfied. The unsatisfied fractions also contribute a small amount to the MIQCP objective. To ensure that cycles and walks have their nodes connected, alternating-edge structure, we must satisfy several auxiliary constraints, enumerated below, with details in the Supplemental Methods (Equations A4.6-A4.23).

1. Each W_i should form a valid walk of alternating sequence and breakpoint (i.e., concordant or discordant) edges.
2. The total CN of all cycles/walks passing through an edge $(u, v) \in E$ is at most $C_l(u, v)$.
3. We require that each cycle/walk traverses through a discordant edge (u, v) at most $R(u, v)$ times. By default the value of $R(u, v)$ is estimated for each discordant edge $(u, v) \in E_d$ based on the number of (long) reads supporting that edge. See Supplemental Methods S4 for details.
4. Each walk W_i (if $z_i > 0$), either forms a cycle starting at node $v_1 \neq s$, or starts at s and ends at t. If W_i forms a cycle we require that the concordant or discordant edge connected to v_1 occurs only once in the cycle.
5. x_{uvi} and z_i are consistent. $z_i = 1 \Leftrightarrow x_{uvi} > 0$ for some $(u, v) \in E$.
6. **Connectivity.** We use auxiliary variables to encode the "discovery order" of the nodes in walk W_i. These variables number the nodes from '1' for the start node, and incrementing by one for each subsequent node in the cycle/walk.
7. **Subwalk constraints.** We enforce a weak constraint by requiring each walk $p_j \in \mathcal{P}$ to be present as a subgraph of the graph induced by some walk W_i.

MIQCP-greedy(α, β, γ, □). For a large graph (e.g. $|E| > 100$), MIQCP-full could be resource intensive. Therefore, we also implemented an MIQCP with additional parameters γ, □, but not k, that identifies only a single walk maximizing the copy number, and additional subwalk constraints satisfied, with parameter γ controlling the weight of the two objectives. Let

$$\bar{\mathcal{P}} = \{j \mid \text{path } p_j \text{ is not satisfied by any previously selected walk}\}.$$

Then the greedy MIQCP objective to identify the next walk W_i is given by:

$$\max \sum_{(u,v)\in W_i \cap E_s} w \cdot x_{uv} \cdot l(u,v) + \gamma \cdot \sum_{j \in \bar{\mathcal{P}}} P_j \qquad (7)$$

Each time a new walk is computed, its copy number is removed for all edges it passed through, and $\bar{\mathcal{P}}$ is updated. The procedure is repeated until either $\alpha \cdot C_l(\mathcal{G})$ copy numbers and $\beta \cdot m$ subwalk constraints are explained by the currently selected walks, or the copy number of next walk is less than $\epsilon \cdot C_l(\mathcal{G})$, for parameter □. We empirically set $\gamma = 0.01 C_l(\bar{\mathcal{G}})/|\bar{\mathcal{P}}|$, where $\bar{\mathcal{G}}$ denote the remaining length-weighted copy number of $\mathcal{G}$ after removing the copy numbers from the last walk, and □ = 0.005. The greedy MIQCP is solved using the same set of auxiliary constraints as before.

Implementation details. In practice, if $\mathcal{G}$ has $|E| > 100$ edges, we use the iterative greedy MIQCP, until either 90% of CN weight is removed from the graph, or the CN weight of the next cycle is less than 1% of the total CN weight. Otherwise, we run full-MIQCP with $\alpha = 0.9, \beta = 0.9$. Initially, $k = 10$, and it is doubled until a feasible solution is reached. If doubling the number of cycles/paths leads to more than 10000 variables in the integer program, we switch to greedy-MIQCP. CoRAL provides users an option to postprocess the greedy-MIQCP solutions with full MIQCP with $\alpha = \min(0.9, 1 - \mathrm{C}_l(\bar{\mathcal{G}})/\mathrm{C}_l(\mathcal{G})), \beta = \min(0.9, 1 - |\bar{\mathcal{P}}|/|\mathcal{P}|)$.

Data Access

All raw and processed sequencing data generated in this study have been deposited to the NCBI BioProject database under accession number PRJNA1110283 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1110283). Other short read and long read sequencing data used in this study can be found in Supplemental Methods S6.

The version of CoRAL used in the presented analysis is included as Supplemental Code. An up-to-date version of CoRAL is available on GitHub https://github.com/ AmpliconSuite/CoRAL.

Competing Interest Statement

V.B. is a co-founder, paid consultant, SAB member and has equity interest in Boundless Bio, inc. and Abterra, Inc. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict-of-interest policies. M.G.J. consults for and holds equity in Vevo Therapeutics. H.Y.C. is a co-founder of Accent Therapeutics, Boundless Bio, Cartography Biosciences, Orbital Therapeutics, and an advisor of 10x Genomics, Arsenal Biosciences, Chroma Medicine, and Spring Discovery. P.S.M. is a co-founder and advisor of Boundless Bio. The remaining authors declare no competing interests.

Acknowledgements

We thank members of the Bafna, Chang, and Mischel labs as well as members of eDyNAmiC for helpful discussion and feedback. We thank Mădălina Giurgiu and the Henssen lab for helpful discussion around running and interpreting results from Decoil. We thank Suhas Srinivasan for technical assistance in cloud computing.

This work was delivered as part of the eDyNAmiC team supported by the Cancer Grand Challenges partnership funded by Cancer Research UK (CGCATF-2021/100012 [P.S.M., H.Y.C.]; CGCATF-2021/100025 [V.B.]) and the National Cancer Institute (OT2CA278688 [P.S.M., H.Y.C.]; OT2CA278635 [V.B.]); U24CA264379, R01GM114362 (V.B.); and National Cancer Institute of the National Institutes of Health under Award Number K99CA286968 (M.G.J.). K.L.H. was supported by a Stanford Graduate Fellowship and an NCI Predoctoral to Postdoctoral Fellow Transition Award (NIH F99CA274692). H.Y. is a Howard Hughes Medical Institute Awardee of the Life Sciences Research Foundation. H.Y.C. is an Investigator of the Howard Hughes Medical Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions

K.Z., M.G.J., J.L., and V.B. conceived the project. K.Z. and V.B. conceived of the CoRAL algorithm with input from M.G.J. and J.L. and assumptions and optimizations. K.Z. developed the CoRAL software with assistance from M.G.J. and J.L. H.Y., I.T.L.W., and S.Z. provided monoclonalized cell lines. K.L.H. performed Nanopore sequencing data for GBM39 and GBM39-HSR cell lines. M.G.J. performed Nanopore and Illumina sequencing GBM39 (mono), COLO320-DM (mono), PC3-DM (mono), and PC3-HSR (mono) cell lines. M.G.J. performed all simulations using ecSimulator and implemented benchmarking pipelines for AmpliconArchitect, CoRAL, and Decoil, with input from J.L. and K.Z. K.Z. performed reconstruction of amplicons using CoRAL and AmpliconArchitect on all cell line data and analyzed results. X.B., K.Z., and J.L. developed software for visualizing reconstructed amplicons. V.B., P.S.M., and H.Y.C. guided data analysis, provided feedback on experimental design, and supervised this work. K.Z., M.G.J., J.L., and V.B. wrote the manuscript with input from all authors.

References

Alekseyev MA, Pevzner PA. 2009. Breakpoint graphs and ancestral genome reconstructions. *Genome res* **19**: 943-957.

Bafna V, Mischel PS. 2022. Extrachromosomal DNA in cancer. *Annu Rev Genom Hum Genet* **23**: 29-52.

Bafna V, Pevzner PA. 1996. Genome rearrangements and sorting by reversals. *SIAM J Comput*

25: 272-289.

Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, et al. 2010. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**: 899-905.

Cao H, Hastie AR, Cao D, Lam ET, Sun Y, Huang H, Liu X, Lin L, Andrews W, Chan S, et al. 2014. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *GigaScience* **3**: 2047-217X.

Chapman OS, Luebeck J, Sridhar S, Wong ITL, Dixit D, Wang S, Prasad G, Rajkumar U, Pagadala MS, Larson JD, et al. 2023. Circular extrachromosomal DNA promotes tumor heterogeneity in high-risk medulloblastoma. *Nat Genet* **55**: 2189-2199.

Deshpande V, Luebeck J, Nguyen NP, Bakhtiari M, Turner KM, Schwab R, Carter H, Mischel PS, Bafna V. 2019. Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat Commun* **10**: 392.

Giurgiu M, Wittstruck N, Rodriguez-Fos E, González RC, Brückner L, Krienelke-Szymansky A, Helmsauer K, Hartebrodt A, Euskirchen P, Koche RP, Haase K, et al. 2023. Decoil: Reconstructing extrachromosomal DNA structural heterogeneity from long-read sequencing data. bioRxiv doi:10.1101/2023.11.15.567169

Hadi K, Yao X, Behr JM, 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.

Helmsauer K, Valieva ME, Ali S, Chamorro González R, Schöpflin R, Röefzaad C, Bei Y, Dorado Garcia H, Rodriguez-Fos E, Puiggròs M, et al. 2020. Enhancer hijacking determines extrachromosomal circular *MYCN* amplicon architecture in neuroblastoma. *Nat Commun* **11**: 5823.

Holtgrewe M. 2010. Mason–a read simulator for second generation sequencing data. *Technical Report FU Berlin*. URL: http://publications.imp.fu-berlin.de/962/

Hung KL, Yost KE, Xie L, Shi Q, Helmsauer K, Luebeck J, Schöpflin R, Lange JT, Chamorro González R, Weiser NE, et al. 2021. ecDNA hubs drive cooperative intermolecular oncogene expression. *Nature* **600**: 731-736.

Hung KL, Luebeck J, Dehkordi SR, Colón CI, Li R, Wong ITL, Coruh C, Dharanipragada P, Lomeli SH, Weiser NE, et al. 2022. Targeted profiling of human extrachromosomal DNA by

CRISPR-CATCH. *Nat Genet* **54**: 1746-1754.

Kim H, Nguyen NP, Turner K, Wu S, Gujar AD, 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.

Lange JT, Rose JC, Chen CY, Pichugin Y, Xie L, Tang J, Hung KL, Yost KE, Shi Q, Erb ML, et al. 2022. The evolutionary dynamics of extrachromosomal DNA in human cancers. *Nat Genet* **54**: 1527-1533.

Lin Y, Nurk S, Pevzner PA. 2014. What is the difference between the breakpoint graph and the de Bruijn graph? *BMC Genom* **15**: S6. doi: 10.1186/1471-2164-15-S6-S6

Luebeck J, Coruh C, Dehkordi SR, Lange JT, Turner KM, Deshpande V, Pai DA, Zhang C, Rajkumar U, Law JA, et al. 2020. AmpliconReconstructor integrates NGS and optical mapping to resolve the complex structures of focal amplifications. *Nat Commun* **11**: 4374.

Luebeck J, Ng AWT, Galipeau PC, Li X, Sanchez CA, Katz-Summercorn AC, Kim H, Jammula S, He Y, Lippman SM, et al. 2023. Extrachromosomal DNA in the cancerous transformation of Barrett's oesophagus. *Nature* **616**: 798-805.

Ly P, Teitz LS, Kim DH, Shoshani O, Skaletsky H, Fachinetti D, Page DC, Cleveland DW. 2017. Selective Y centromere inactivation triggers chromosome shattering in micronuclei and repair by non-homologous end joining. *Nat Cell Bio* **19**: 68-75.

Nathanson DA, Gini B, Mottahedeh J, Visnyei K, Koga T, Gomez G, Eskin A, Hwang K, Wang J, Masui K, Paucar A. 2014. Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant *EGFR* DNA. *Science* **343**: 72-76.

Passananti C, Davies B, Ford M, Fried M. 1987. Structure of an inverted duplication formed as a first step in a gene amplification event: implications for a model of gene amplification. *EMBO J* **6**: 1697-1703.

Prensner JR, Chen W, Han S, Iyer MK, Cao Q, Kothari V, Evans JR, Knudsen KE, Paulsen MT, Ljungman M, et al. 2014a. The long non-coding RNA *PCAT-1* promotes prostate cancer cell proliferation through *cMyc*. *Neoplasia* **16**: 900-908.

Prensner JR, Chen W, Iyer MK, Cao Q, Ma T, Han S, Sahu A, Malik R, Wilder-Romans K, Navone N, et al. 2014b. *PCAT-1*, a long noncoding RNA, regulates *BRCA2* and controls homologous recombination in cancer. *Cancer Res* **74**: 1651-1660.

Shoshani O, Brunner SF, Yaeger R, Ly P, Nechemia-Arbely Y, Kim DH, Fang R, Castillon GA, Yu M, Li JS, et al. 2021. Chromothripsis drives the evolution of gene amplification in cancer. *Nature* **591**: 137-141.

Steele CD, Abbasi A, Islam SA, Bowes AL, Khandekar A, Haase K, Hames-Fathi S, Ajayi D, Verfaillie A, Dhami P, et al. 2022. Signatures of copy number alterations in human cancer. *Nature* **606**: 984-991.

Talevich E, Shain AH, Botton T, Bastian BC. 2016. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput Biol* **12**: e1004873.

Turner KM, Deshpande V, Beyter D, Koga T, Rusert J, Lee C, Li B, Arden K, Ren B, Nathanson DA, et al. 2017. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* **543**: 122-125.

Wu S, Turner KM, Nguyen N, Raviram R, Erb M, Santini J, Luebeck J, Rajkumar U, Diao Y, Li B, et al. 2019. Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature* **575**: 699-703.

Xiong T, Li J, Chen F, Zhang F. 2019. *PCAT-1*: a novel oncogenic long non-coding RNA in human cancers. *Int J Biol Sci* **15**: 847.

Yang C, Chu J, Warren RL, Birol I. 2017. NanoSim: nanopore sequence read simulator based on statistical characterization. *GigaScience* **6**: gix010.

Zhu Y, Gujar AD, Wong CH, Tjong H, Ngan CY, Gong L, Chen YA, Kim H, Liu J, Li M, et al. 2021. Oncogenic extrachromosomal DNA functions as mobile enhancers to globally amplify chromosomal transcription. *Cancer Cell* **39**: 694-707.

Figure Legends

Figure 1: Long read based ecDNA reconstruction. (A) Native ecDNA structure, and copy number. (B) Cartoon of the breakpoint graph derived from the ecDNA in (A). Sequence edges represent segments of the reference genome. Concordant edges connect consecutive sequences with respect to the reference genome order, and discordant edges connect non-consecutive genome segments. Nodes are created at the endpoints of each sequence edge, and include source and sink nodes, *s* and *t*. (C) Multiple collections of decomposed paths and cycles from the breakpoint graph explain the changes in copy number and observed SVs. Long-reads that span regions of high multiplicity can help resolve the correct cycle. (D) Overview of the CoRAL method.

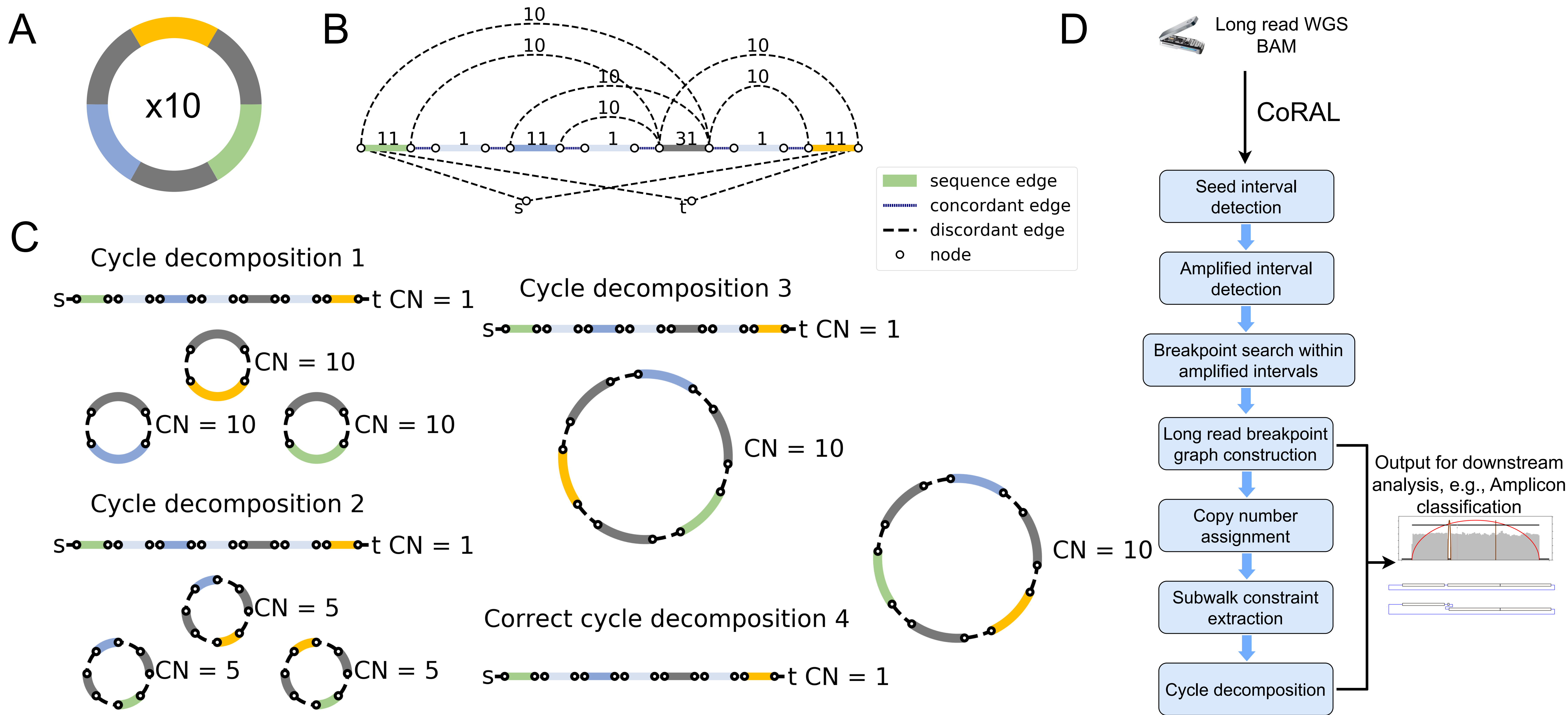
Figure 2: Overview of simulation benchmarking. (A) True structure compared to CoRAL, Decoil, and AA reconstructions for an example amplicon (Episomal, 8 observed breakpoints). (B-E) Cumulative distributions of CoRAL, Decoil, and AA reconstructions across all simulations for (B) breakpoint graph accuracy, (C) cycle interval overlap, (D) cyclic longest common subsequence and (E) rank-order distribution of reconstruction length error. Empirical cumulative densities are reported for (B), (C) and (D); and each point in (E) corresponds to a simulated amplicon. See Supplemental Methods S7 for more detailed information.

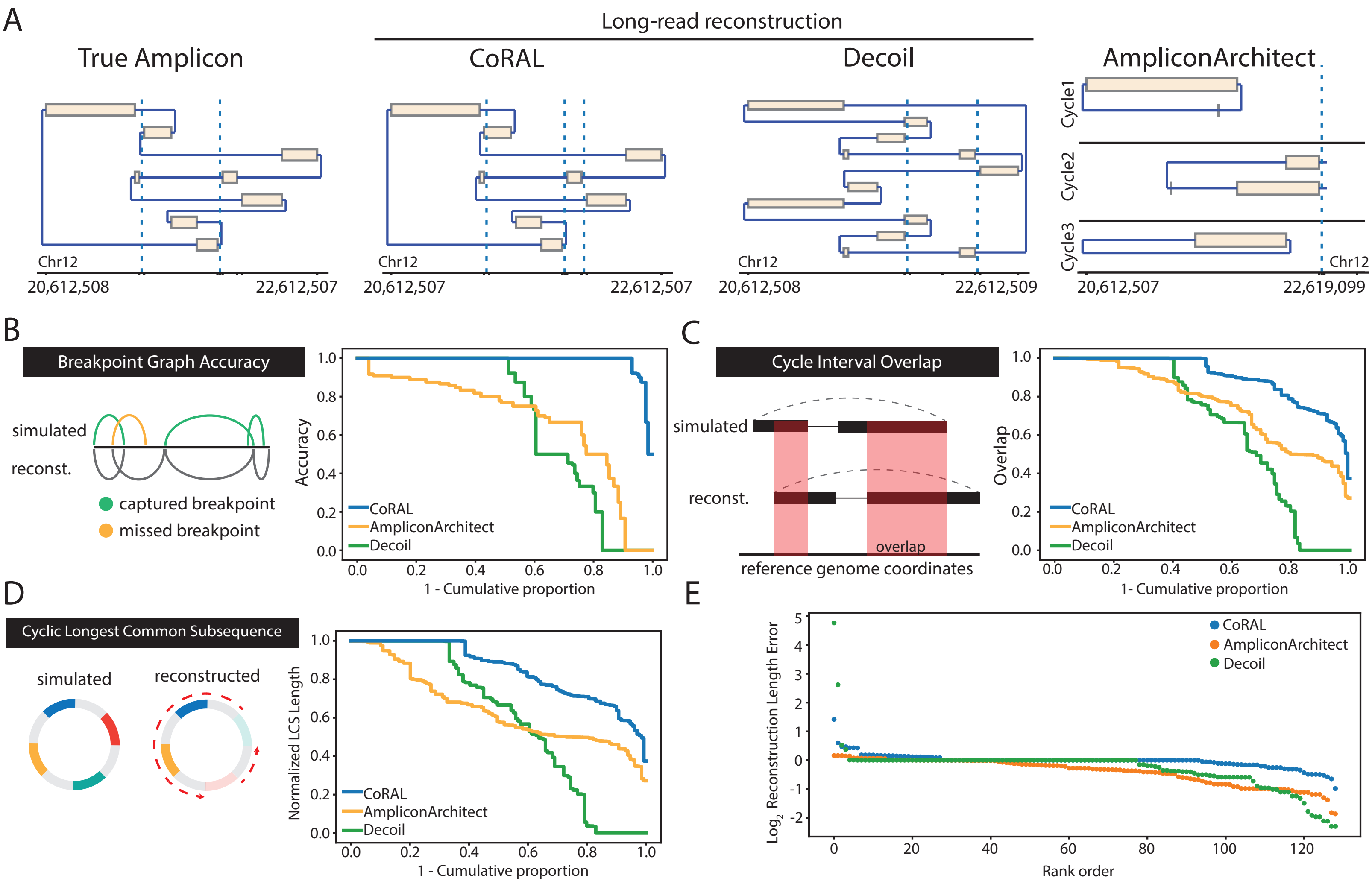
Figure 3: Amplicon reconstruction in cell lines. (A) Fraction of length-weighted copy numbers given by the 3 heaviest cycles, reported by CoRAL and AA; (B) number of satisfied subwalk constraints by CoRAL and AA, in cell lines. (C) The cycle with largest length-weighted CN from previously published COLO320-DM. (D) The cycle with largest length-weighted CN from monoclonalized COLO320-DM. In (C) and (D), each black arc within the cycle indicates a subwalk constraint satisfied by that cycle. (E) The two cycles with largest length weighted copy numbers by CoRAL. (F) The three cycles with largest length weighted copy numbers by AA.

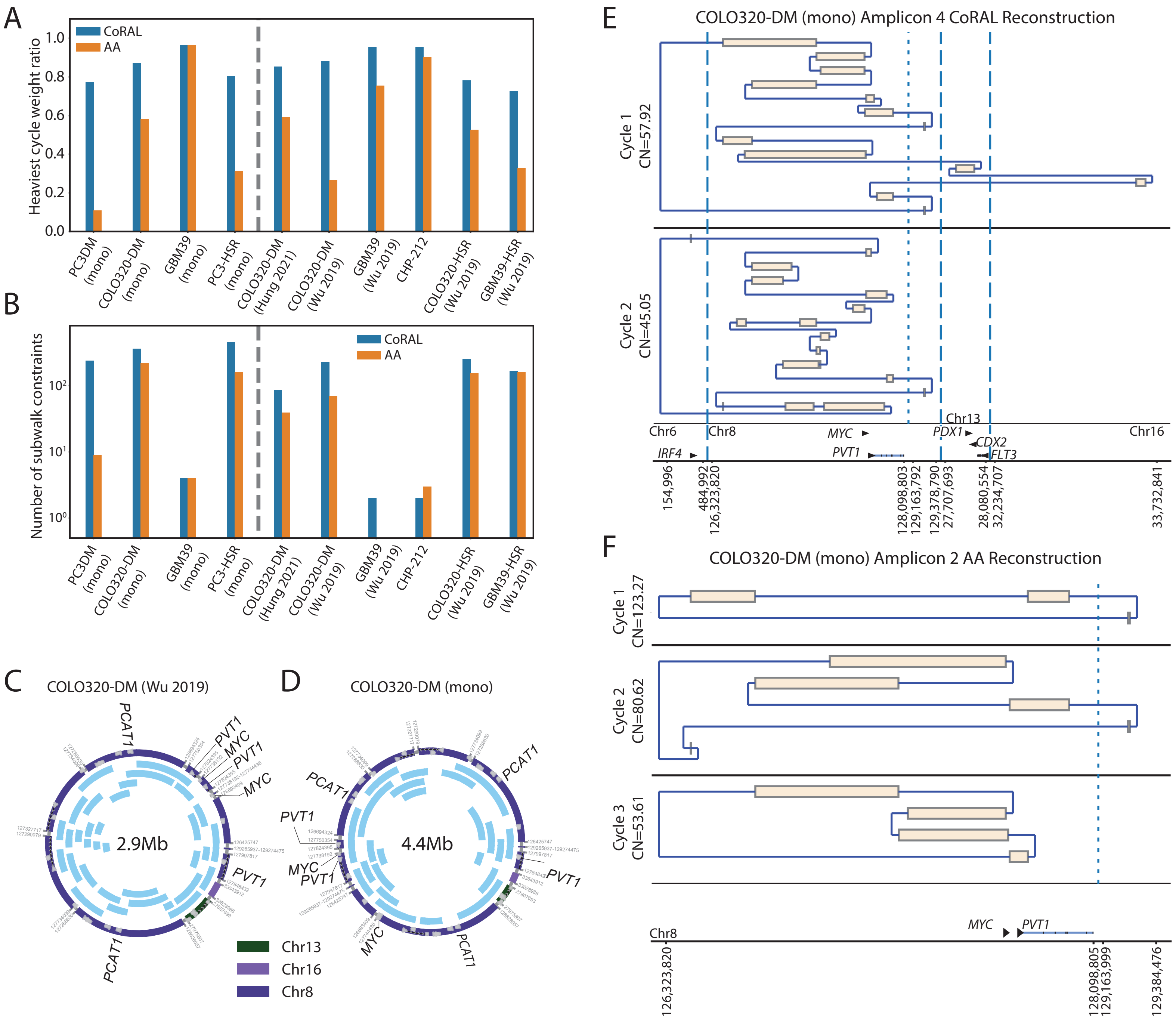
Tables

Cell line	Cancer type	Gene(s)	Amp.	Mono.	Source
PC3DM (mono)	Prostate	*MYC*	ecDNA	Yes	This study
PC3HSR (mono)	Prostate	*MYC*	HSR	Yes	This study
GBM39 (mono)	Glioblastoma	*EGFRvIII*	ecDNA	Yes	This study
COLO320-DM (mono)	Colorectal	*MYC*	ecDNA	Yes	This study
COLO320-DM	Colorectal	*MYC*	ecDNA	No	Hung 2021 (Hung et al. 2021)
COLO320-DM	Colorectal	*MYC*	ecDNA	No	Wu 2019 (Wu et al. 2019)
GBM39	Glioblastoma	*EGFRvIII*	ecDNA	No	Wu 2019 (Wu et al. 2019)
CHP-212	Neuroblastoma	*MYCN*, *TRIB2*	ecDNA	No	Helmsauer 2020 (Helmsauer et al. 2020)
COLO320-HSR	Colorectal	*MYC*	HSR	No	Wu 2019 (Wu et al. 2019)
GBM39-HSR	Glioblastoma	*EGFRvIII*	HSR	No	Wu 2019 (Wu et al. 2019)

Table 1: **Overview of cell lines profiled in this study.** *Cell type name, cancer type, subset of important amplified oncogenes, amplification type, monoclonal status, and source. Abbreviations: HSR = Homogeneously Staining Region; ecDNA = extrachromosomal DNA.*


A
x10
B
10
10
10
10
10
10
11
1
11
1
31
1
11
s
t
sequence edge
concordant edge
discordant edge
node
C
Cycle decomposition 1
s
t CN = 1
CN = 10
CN = 10
CN = 10
Cycle decomposition 2
s
t CN = 1
CN = 5
CN = 5
CN = 5
Cycle decomposition 3
s
t CN = 1
CN = 10
Correct cycle decomposition 4
s
t CN = 1
CN = 10
D
Long read WGS BAM
CoRAL
Seed interval detection
Amplified interval detection
Breakpoint search within amplified intervals
Long read breakpoint graph construction
Copy number assignment
Subwalk constraint extraction
Cycle decomposition
Output for downstream analysis, e.g., Amplicon classification


A
Long-read reconstruction
True Amplicon
CoRAL
Decoil
AmpliconArchitect
Cycle1
Cycle2
Cycle3
Chr12
20,612,508
22,612,507
20,612,507
22,612,507
20,612,508
22,612,509
20,612,507
22,619,099
B
Breakpoint Graph Accuracy
simulated
reconst.
captured breakpoint
missed breakpoint
Accuracy
1 - Cumulative proportion
CoRAL
AmpliconArchitect
Decoil
C
Cycle Interval Overlap
simulated
reconst.
overlap
reference genome coordinates
Overlap
1- Cumulative proportion
D
Cyclic Longest Common Subsequence
simulated
reconstructed
Normalized LCS Length
1 - Cumulative proportion
E
Log$_2$ Reconstruction Length Error
Rank order
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
3
4
5
0
20
40
60
80
100
120


A
Heaviest cycle weight ratio
1.0
0.8
0.6
0.4
0.2
0.0
CoRAL
AA
PC3DM (mono)
COLO320-DM (mono)
GBM39 (mono)
PC3-HSR (mono)
COLO320-DM (Hung 2021)
COLO320-DM (Wu 2019)
GBM39 (Wu 2019)
CHP-212
COLO320-HSR (Wu 2019)
GBM39-HSR (Wu 2019)
B
Number of subwalk constraints
10^2
10^1
10^0
CoRAL
AA
PC3DM (mono)
COLO320-DM (mono)
GBM39 (mono)
PC3-HSR (mono)
COLO320-DM (Hung 2021)
COLO320-DM (Wu 2019)
GBM39 (Wu 2019)
CHP-212
COLO320-HSR (Wu 2019)
GBM39-HSR (Wu 2019)
C
COLO320-DM (Wu 2019)
PCAT1
PVT1
MYC
PVT1
MYC
127288630
127734099
126694324
127750354
127824395
127738192
127824395
127738192-127744436
126693409
127327717
127290079
2.9Mb
126425747
129265937-129274475
127997817
PVT1
127848432
33543912
33628986
27807693
27975807
126626057
127734099
127288630
PCAT1
D
COLO320-DM (mono)
127290079
127327717
127734099
127288630
127734099
127288630
PCAT1
PCAT1
PVT1
126694324
127750354
127824395
127738192
MYC
PVT1
4.4Mb
126425747
129265937-129274475
127997817
127848432
PVT1
33543912
33628986
27807693
27975807
126626057
127997817
129265937-129274475
126425747
126693409
127744436
MYC
PCAT1
Chr13
Chr16
Chr8
E
COLO320-DM (mono) Amplicon 4 CoRAL Reconstruction
Cycle 1
CN=57.92
Cycle 2
CN=45.05
Chr6
Chr8
Chr13
Chr16
MYC
PDX1
CDX2
IRF4
PVT1
FLT3
154,996
484,992
126,323,820
128,098,803
129,163,792
129,378,790
27,707,693
28,080,554
32,234,707
33,732,841
F
COLO320-DM (mono) Amplicon 2 AA Reconstruction
Cycle 1
CN=123.27
Cycle 2
CN=80.62
Cycle 3
CN=53.61
Chr8
MYC
PVT1
126,323,820
128,098,805
129,163,999
129,384,476

CSH PRESS GENOME RESEARCH

CoRAL accurately resolves extrachromosomal DNA genome structures with long-read sequencing

Kaiyuan Zhu, Matthew Gregory Jones, Jens Luebeck, et al.

Genome Res. published online July 9, 2024
Access the most recent version at doi:10.1101/gr.279131.124

Published by Cold Spring Harbor Laboratory Press