AmpliconReconstructor: Integrated analysis of NGS and optical mapping resolves the complex structures of focal amplifications in cancer

Authors: Jens Luebeck[1,2], Ceyda Coruh[3], Siavash R. Dehkordi[2], Joshua T. Lange[4,5], Kristen M. Turner[5], Viraj Deshpande[2], Dave A. Pai[6], Chao Zhang[1], Utkrisht Rajkumar[2], Julie A. Law[3], Paul S. Mischel[5,7,8], Vineet Bafna[2]*

Affiliations:

[1]Bioinformatics and Systems Biology Graduate Program, University of California at San Diego, La Jolla, CA 92093, USA

[2]Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA 92093, USA

[3]Plant Molecular and Cellular Biology Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037, USA

[4]Biomedical Sciences Graduate Program, University of California at San Diego, La Jolla, CA 92093, USA

[5]Ludwig Institute for Cancer Research, University of California at San Diego, La Jolla, CA 92093, USA

[6]Bionano Genomics, Inc., San Diego, CA 92121, USA

[7]Moores Cancer Center, University of California at San Diego, La Jolla, CA 92093, USA

[8]Department of Pathology, University of California at San Diego, La Jolla, CA 92093 USA

*To whom correspondence should be addressed (vbafna [at] cs.ucsd.edu)

Abstract:

Oncogene amplification, a major driver of cancer pathogenicity, is often mediated through focal amplification of genomic segments. Recent results implicate extrachromosomal DNA (ecDNA) as the primary mechanism driving focal copy number amplification (fCNA) - enabling gene amplification, rapid tumor evolution, and the rewiring of regulatory circuitry. Resolving an fCNA's structure is a first step in deciphering the mechanisms of its genesis and the subsequent biological consequences. Here, we introduce a powerful new computational method, AmpliconReconstructor (AR), for integrating optical mapping (OM) of long DNA fragments (>150kb) with next-generation sequencing (NGS) to resolve fCNAs at single-nucleotide resolution. AR uses an NGS-derived breakpoint graph alongside OM scaffolds to produce high-fidelity reconstructions. After validating performance by extensive simulations, we used AR to reconstruct fCNAs in seven cancer cell lines to reveal the complex architecture of ecDNA, breakage-fusion-bridge cycles, and other complex rearrangements. By distinguishing between chromosomal and extrachromosomal origins, and by reconstructing the rearrangement signatures associated with a given fCNA's generative mechanism, AR enables a more thorough understanding of the origins of fCNAs, and their functional consequences.

Main:

Oncogene amplification is a major driver of cancer pathogenicity[1–5]. Genomic signatures of oncogene amplification include somatic focal Copy Number Amplifications (fCNAs) of small (typically < 10Mbp) genomic regions[5,6]. Multiple mechanisms cause fCNAs including, but not limited to, extrachromosomal DNA (ecDNA) formation[5,7,8], chromothripsis[9], tandem duplications[10,11] and breakage-fusion-bridge (BFB) cycles[12–14]. ecDNA, in particular, enables tumors to achieve far higher oncogene genomic copy numbers and maintain far greater levels of intratumor genetic heterogeneity than previously anticipated, due to their non-chromosomal mechanism of inheritance - enabling tumors to evolve rapidly[5,15,16]. In addition, the very high DNA template level generated by ecDNA-based amplification, coupled to its highly accessible chromatin architecture, permits massive oncogene transcription[17–19].

While ecDNA elements are a common form of fCNA[5], other mechanisms can also result in amplification with very different functional consequences[6]. Thus, accurate identification and reconstruction of the fCNA structure not only describes the rearranged genomic landscape, but also represents a first step in identifying the generative mechanism. Reconstruction of fCNA architecture involves determining the order and orientation of the genomic segments that constitute the amplicon. There are many methods to detect single genomic breakpoints from sequencing data using a variety of different sequencing technologies[20–23]. However, fewer methods are available to handle the more difficult problem of ordering and orienting multiple genomic segments joined by breakpoints into high confidence scaffolds which enable complete reconstructions of complex rearrangements[6,24].

A previous method for characterizing the identity of focally amplified genomic regions, AmpliconArchitect (AA), generates an accurate breakpoint graph from next-generation sequencing (NGS) data[6]. The graph encodes the genomic segments involved in fCNAs, their copy numbers, and breakpoint edges connecting the segments. Unambiguous reconstruction of fCNA architecture requires extracting paths and cycles from the breakpoint graph, to reveal the true structure of the underlying rearranged genome. However, in practice, path/cycle extraction is often confounded by duplications of large genomic regions inside an amplicon (Supplemental Fig. S1a), imperfections in the graph arising from errors in estimation of segment copy numbers, or erroneous and/or missing breakpoints.

We hypothesized that an approach combining the strengths of NGS with long-range mapping data would enable larger and more unambiguous reconstructions of fCNA architectures. To produce the highest-quality reconstructions of large, complex focal amplifications, we utilized both optical mapping (OM) data as well as the breakpoint graphs generated from AA with NGS data. OM provides single-molecule information about the approximate locations of fluorescently-labeled sequence motifs on long fragments of DNA[25]. The median molecule ("map") length used in assembly across all samples used in this study is 244 kbp (molecule N50 340 kbp), while the median segment length in breakpoint graphs used in this study is 100 kbp, highlighting that OM data can span multiple junctions in breakpoint graphs derived from focal amplifications (Supplemental Table 1). The integrated NGS data and OM data provide an orthogonal pairing of short- and long-range information about genomic structural variation. We utilized Bionano (Bionano Genomics, Inc., San Diego, CA) whole-genome imaging to generate single-molecule optical maps, which were subsequently *de novo* assembled into OM contigs (contig N50 72.8 Mbp) to improve confidence and reduce noise.

Here we present a novel computational method for reconstructing large complex fCNAs, AmpliconReconstructor (AR). AR takes a breakpoint graph and long-range Bionano OM data as inputs. AR produces an ordering and orientations of graph segments, with fine-structure information from the breakpoint graph embedded into the large-scale reconstructions. AR outputs megabase-scale reconstructions of fCNA amplicons. We demonstrated the fine-scale accuracy of AR using simulated OM data derived from previously analyzed cancer cell lines[6,21]. Furthermore, we reconstructed fCNAs at unprecedented resolution and size in seven cancer cell lines (CAKI2, GBM39, NCI-H460, HCC827, HK301, K562, T47D). Finally, we validated the reconstructions using cytogenetics.

Results:

AR separates the computational tasks involved in reconstruction of fCNAs into three primary modules (Fig. 1a,b). First, an OM alignment module, SegAligner, for aligning reference segments to assembled OM contigs generated by either the Bionano Irys or Bionano Saphyr instruments (Supplemental Fig. S1b-d, Methods - “Optical map contig alignments with SegAligner”). SegAligner is critical to the imputation process as it can score placements of short genomic segments onto an OM contig, which wasn’t possible with other aligners. Second, a scaffolding module, which takes a collection of breakpoint graph segments aligned to OM contigs as input and creates scaffolds represented by directed acyclic graphs (DAGs) (Fig. 1c-e, Methods – “Reconstructing amplicon paths with AmpliconReconstructor”). Third, AR also relies on a novel scaffold-path imputation technique (Fig. 1f-h, Methods – “Imputing paths in the scaffold with

AmpliconReconstructor") to connect and chain together breakpoint graph segments that may individually be too small to be informatively labeled and aligned with optical mapping (Fig. 1f). Finally, a pathfinding module, which links scaffolds and searches for paths in a copy number (CN)-aware manner, to identify possible reconstructions of the amplicon. AR outputs a collection of sequence resolved paths supported by the linked scaffolds. To visualize output from AR, we implemented a visualization utility, CycleViz, to show the integrated OM- and NGS-derived breakpoint graph data (Supplemental Fig. S2). AmpliconReconstructor is implemented in Python 2.7, and SegAligner is implemented in C++. Both tools are available publicly at https://github.com/jluebeck/AmpliconReconstructor.

AR accurately reconstructs simulated amplicons

We utilized multiple simulation strategies to measure the performance of AR. For a ground-truth set of amplicon structures, we used 85 non-trivial amplicon breakpoint graph paths previously reported by AmpliconArchitect from 25 cancer cell lines[6]. The breakpoint graph paths included both cyclic and non-cyclic paths with lengths varying from 260 kbp to 2.8 Mbp (median 1.1 Mbp) and the number of graph segments varying from 3 to 47 (mean 17.5 segments; Supplemental Table 2). These paths were used as a reference to simulate OM molecules. (Methods – "Simulation of amplicons to measure AR performance"). Simulated molecules were assembled into contigs using the Bionano Assembler[26,27].

For each of the 85 simulation cases, we then ran AR on the corresponding breakpoint graph and the *de novo* assembled contigs, and examined four different variables that could affect the performance of AR. First, we tested AR performance using SegAligner for OM alignment, versus AR using other OM alignment tools to replace SegAligner. Second, we evaluated the performance of AR across a range of amplicon copy numbers. Third, we measured performance with false edges present in the breakpoint graph. Finally, we generated and tested mixtures of three similar amplicons from the same samples, simulated with different amplicon copy numbers, to measure the effects of potential amplicon heterogeneity on AR performance.

We measured the accuracy of AR by computing precision and recall across the four simulation conditions. As precision and recall could be quantified in multiple ways when comparing ground-truth and reconstructed simulation paths, leading to different understandings of performance, we described three ways of measuring the similarity of the paths (“Length (bp)”, “Nseg”, “Breakpoint”; Methods – “Measuring AR simulation performance”), based on the longest common substring (LCS) between ground-truth and reconstructed path sequences. We report the “Length (bp)” measurement in the analysis described here, while results with other measurements are presented in Supplemental Table 2 and Supplemental Figure S3.

AR using SegAligner achieved a mean F1 score (harmonic mean of the precision and recall) of 0.88 for the highest copy number level (CN 20) and 0.68 for the lowest copy number level (CN 2) (Fig. 1i, Supplemental Fig. S3, Supplemental Table 2). In contrast, when OMBlast[28] or Bionano RefAligner[26,29] were used in place of SegAligner, we noticed a decrease in both precision and recall. For RefAligner and OMBlast, respectively, we report mean F1 scores of 0.52, 0.43 for CN 20, and 0.42, 0.41 for CN 2. When imputation was omitted from AR, the mean F1 score for CN 20 decreased from 0.88 to 0.70. We observed similarly consistent trends using other methods of measuring precision and recall – “Nseg” and “Breakpoint” (Supplemental Fig. S3). We saw a few cases of ‘assembly failure,’ where no paths differing from the reference genome involving the amplicon segments were assembled. Figure 1i shows cumulative precision and recall values for AR using SegAligner (with and without imputation), and with assembly failures filtered. We additionally reported simulation F1 scores with and without filtering for possible OM assembly failure (Supplemental Table 2).

False edges in the breakpoint graph increase the possible number of path imputations that AR must consider, potentially leading to erroneous scaffolds. We designed another simulation study where after simulating CN 20 amplicon OM data, additional false edges were added between existing graph segments. We tested three scenarios with the proportion of additional false edges ranging from 0%, 50% and 100% of the number of true graph edges. The three scenarios resulted in nearly identical mean F1 scores of 0.881, 0.880, 0.881 across the 85 amplicon simulations (Supplemental Table 2, Supplemental Fig. S4a), highlighting the robustness of the path imputation method.

To understand how AR performed when faced with structural heterogeneity, we designed a simulation study involving 123 combinations of breakpoint graph paths where each combination was derived from a single sample at varying copy number mixtures. We simulated amplicons from heterogeneous mixtures with (1) a single dominant amplicon (CNs 20-2-2); (2) a linear mixture of CNs (CNs 20-15-10); (3) equally abundant amplicons (CNs 20-20-20). We report mean F1 scores of 0.92, 0.89, and 0.91, respectively for the three cases (Supplemental Table 2). To explain the increase in performance of the mixture simulations as compared to the single amplicon simulations, we hypothesize that the greater total number of molecules improved the assembly process. Regardless, the high similarity between the precision and recall in each mixture case (Supplemental Fig. S4b) indicates AR can reconstruct an accurate amplicon path even in the context of heterogeneity. Based on these metrics, we found AR to be robust, and to outperform other methods. To further demonstrate its ability to reconstruct a variety of complex fCNAs, we ran AR on seven cancer cell lines with evidence of fCNA.

AR reconstructs ecDNA in multiple forms

Three cell lines in our data set were previously reported to contain ecDNA[5] - GBM39, NCI-H460, and HK301. In a previous study[17], we analyzed the glioblastoma multiforme (GBM) cell line GBM39 using a preliminary version of AR that used RefAligner and manual merging of graph segments, but without path imputation or scaffold linking capabilities. Re-analysis reproduced an unambiguous 1.26 Mbp EGFRvIII-containing circular ecDNA that was identical to the previously published structure[17] (Supplemental Fig. S5). The entire structure was captured by a single non-circular OM contig, with circularity confirmed by an overlapping graph segment aligned to both ends of the contig.

Previous studies of ecDNA have documented their integration into chromosomes over time, linearizing and appearing as homogeneously staining regions (HSRs), often in nonnative locations[5,7,15]. In a previous study[5], The GBM cell line, HK301, had been cytogenetically determined to have circular ecDNA; however, we observed from FISH (fluorescence in situ hybridization) data that the sample's ecDNA had become HSR-like at the time of this study (Fig. 2a). AA generated a breakpoint graph supporting amplification of both EGFRvIII and EGFR wild-type (Fig. 2c), however an unambiguous reconstruction from the graph alone was not possible. The AR reconstruction of the HK301 fCNA indicated a complex and cyclic structure supported by three contigs (Fig. 2d), which explained 98.1% of the amplified genomic regions. The graph segments came predominantly from chr7, but also included two small regions (2890 bp, 4591 bp) from chr6 (Fig. 2c,d). We noted a ~20 kbp deletion inside EGFR, showing a lower CN than the surrounding region, but which was still amplified over the baseline, non-amplicon regions of chr7. This indicates heterogeneity of EGFR wild-type/vIII mutation status. Despite the heterogenous status of this allele, AR reconstructed the EGFRvIII version – which is the dominant form of the amplicon (Fig. 2d).

The lung cancer cell line NCI-H460 has previously been documented to bear MYC amplification[30], and our cytogenetic analysis showed evidence for both its HSR-like and ecDNA amplification (Fig. 2e,f). Despite the heterogeneous nature of the amplicon's integration status, AA generated a breakpoint graph for a contiguous 2.15 Mbp region of chr8 (Fig. 2g). AR reconstructed a single 4.10 Mbp structure supported by five OM contigs (Fig. 2h). This structure contained all amplified segments from the breakpoint graph and explained the breakpoint graph segment copy number ratios of the duplicated segments. For example, segment chr8:129,404,278-129,591,422 appeared 4 times, chr8:128,690,200-129,404,277 (carrying MYC & PVT1) appeared twice, chr8:129,591,423-129,911,811 appeared twice, and chr8:129,911,812-130,640,594 appeared once, making the ratios consistent with the estimated graph segment copy numbers (46, 25, 25, 12, respectively; Fig. 2g). The status of the long non-coding RNA PVT1 (a known regulator of MYC)[31] on this amplicon is heterogeneous, as one copy of PVT1 does not contain breakpoints, while the other shows a disrupted copy of PVT1. AR also identified a self-inversion at the end of the amplicon (black arrows in Fig. 2h), suggestive of an alternating forward-backward orientation (segmental tandem aggregation with inversion) of the amplicon in the agglomerated ecDNA.

In summary, AR reconstructed paths that were consistent with the expected copy number ratios and graph structures in GBM39, HK301, and NCI-H460, explaining 99.9%, 98.1%, and 100% of the amplified genomic content in the breakpoint graphs for each cell line, respectively. Furthermore, the AR reconstructions of ecDNA in HSR-like form lend additional evidence to the agglomerative model of ecDNA integration (Fig. 2b)[8,32,33].

AR reconstructs a rearranged Philadelphia chromosome in K562

The classical model of the BCR-ABL1 (Philadelphia chromosome) fusion involves a reciprocal translocation of the q arms of chromosomes 9 and 22[34]. However, this mechanism alone does not explain the copy number amplification of BCR-ABL1 fusion commonly observed in chronic myeloid leukemia (CML), highlighting a need for methods to better understand the genesis of the BCR-ABL1 amplification[35,36]. To reconstruct the fine structure of a Philadelphia chromosome, we used the CML cell line K562 where a BCR-ABL1 fusion had previously been reported[37].

The AA reconstructed breakpoint graph of the BCR-ABL1 fCNA in K562 (Fig. 3a) contains 8.5 Mbp of amplified genomic segments. The graph shows signatures of complex rearrangements alongside the BCR-ABL1 fusion, which AA predicted to have a copy number of 17 (Fig. 3a). We generated both Bionano Irys and Bionano Saphyr OM data for K562 cells and observed consistent results in the independent reconstructions of amplicons from both sources (Supplemental Fig. S6a,b). Using the breakpoint graph and OM contigs, AR reconstructed a complex linear structure that chained together 1.7 Mbp from chr22 (containing BCR), 548 kbp of chr9 (containing ABL1), and multiple regions from chr13 (732 kbp; including a disrupted copy of GPC5) (Fig. 3b). In Figure 3b, we show one possible scaffolding of the given regions, whose structure was reproduced in both Saphyr and Irys datasets. AR also reported junctions between segments in the breakpoint graph where NGS-derived breakpoint edges were not reported, as indicated by the missing half-height grey bars between adjacent genomic segments in the genome tracks of Figure 3b.

We performed multiple FISH experiments using combinations of probes for BCR, ABL1, GPC5, and chr22 centromere probe CEP22. The FISH images confirmed the co-localization of the BCR-ABL1 fusion and GPC5 on a common HSR-like structure (Fig. 3c). Furthermore, it validated the status of the K562 BCR-ABL1 fusion as being located on chr22 (Supplemental Fig. S7).

In addition to the reconstruction reported in Figure 3b, AR additionally identified other scaffolds, indicating that the genomic structure surrounding the BCR-ABL1 translocation may be varied across the multiple copies (Supplemental Fig. S6c,d; Supplemental Fig. S8a-f). In particular, the genomic segment bearing CLTCL appears in both forward and reverse directions (Supplemental Fig. S8b,c). Other amplified regions of chr13 include a self-inversion at the 3' end of GPC5 (Supplemental Fig. S6c,d, Supplemental Fig. S8e). A scaffold from the Irys-based reconstruction indicated a secondary reconstruction could be joined with the BCR-ABL1 reconstruction (Supplemental Fig. S6d; overlap of segment 20). From the AR reconstructions of the BCR-ABL1 amplicon and the co-existence of BCR, ABL1 and GPC5 in overlapping locations, as shown by FISH (Fig. 3c 'Zoom'), AR enabled us to hypothesize a potential sequence of events by which the fCNA formed. The AR reconstructions support the formation of the BCR-ABL1 translocation (Supplemental Fig. S8g;i-ii) followed by incorporation of chr13 regions (Supplemental Fig. S8g;iii-iv), which subsequently undergo rearrangement (Supplemental Fig. S8g;v), and ultimately a series of inverted repeats, possibly mediated through dicentrism (Supplemental Fig. S8g;vi).

These results are consistent with previous reports that used cytogenetic approaches in BCR-ABL1-positive samples to identify the presence of additional chromosomal segments besides chr9 and chr22 involved in the Philadelphia chromosome[30,31]. AR reconstructed the first base-pair resolved structures of the surrounding complex rearrangement. The rearrangement of BCR-ABL1 and chr13 segments was followed by additional duplications leading to a focal amplification. This example demonstrates the utility of AR in resolving complex fCNAs, enhancing our understanding of the fundamental mechanisms of cancer pathogenesis.

AR enabled the first sequence-based reconstruction of a breakage-fusion-bridge

The BFB mechanism of genomic amplification involves the loss of telomeres and subsequent fusion of two sister chromatids[12,13]. In subsequent cellular division, the asymmetric breaking of the fused dicentric chromosome structure results in one daughter cell having an increased copy number of pieces of the previously fused chromosome. The structure of various BFBs have been analyzed using cytogenetic techniques[14] and also by computational models that predict a BFB mechanism based on copy number counts[38,39]. Both methods are imprecise, to a degree, and may fail to capture the fine structure of the BFB or handle imprecise copy number counts and/or additional structural variants (SVs) inside the BFB. We deployed AR on the HCC827 lung cancer cell line where we AA and cytogenetics previously suggested a BFB containing EGFR, though an unambiguous structure was not identifiable[5,6].

We observed a banded pattern of EGFR and CEP7 (a chr7 centromeric D7Z1 repeat) in a DNA FISH experiment on HCC827 cells, suggestive of a BFB mechanism (Fig. 4a). AA generated a breakpoint graph of a 4.2 Mbp amplified region of chr7 containing EGFR (Fig. 4b). The amplified BFB segments in the AA output ranged in size from 217 kbp to 1176 kbp. AR enabled the reconstruction of 16 unique OM scaffolds which, when combined, covered the entirety of a BFB structure (Fig. 4c,d). The five most informative single scaffolds ranged in size from 750 kbp to 2.3 Mbp, containing multiple junctions which validate the order and orientation of the BFB breakpoint graph segments, resulting in a 9.4 Mbp BFB structure, hereafter referred to as a BFB repeat unit. The BFB repeat unit was amplified across the chromosome (Fig. 4a, e-f). AR also revealed a region outside the AA amplicon, near the centromere of chr7, which explained the observed EGFR and CEP7 repeat (“F”). In segment “B”, we observed both a 600 bp deletion across the entire BFB repeat unit and an 11 kbp inversion. The latter is labeled throughout Figure

4 with a black asterisk and only appears when segment "B" is duplicated and inverted, suggesting that the SV arose during the formation of the BFB.

When the AR scaffolds were combined with the copy number data present in the breakpoint graph, we identified a single BFB structure, that was consistent with the theoretical BFB model of BFB formation[40]. A putative sequence of BFB cycles and additional structural variation that results in the final BFB structure is shown in Fig. 4f (also Supplemental Fig. S9a,b). Note that the copy number information and the theoretical model together could not have reconstructed this BFB, as it contains heterogeneous interior structural variants. We further validated the BFB patterning in HCC827 cells with multi-FISH for segments "A", "C", and "D" from the BFB, using FISH (Fig. 4e, Supplemental Fig. S9c). Together, these results on HCC827 show the power of AR as a method to elucidate a complex mechanism of BFB-driven fCNA, even in the presence of additional structural variant heterogeneity.

In addition to the EGFR-bearing amplicon, AA detected 5 other amplicons containing MYC and NCOA2, among other oncogenes, in HCC827. The graphs were complex (Supplemental Fig. S10a) and in many cases AA did not identify discordant edges between distinctly amplified regions. Given the dearth of breakpoint edges, we combined the amplicon breakpoint graphs for all six HCC827 amplicons and ran AR on the combined graph, containing 555 segments. AR identified 206 contigs having alignments to one or more graph segments. AR reconstructed multiple possible scaffolds and captured overlapping subsets of amplicon regions from different graphs, suggestive of possible heterogeneity. One scaffold showed NCOA2 located on a native region of chr8, while another showed NCOA2 joined to MYC through a segment of chr21 (Supplemental Fig. S10b,c).

Other focal amplifications reconstructed by AR

In breast cancer cell line T47D, where the AA breakpoint graph suggested amplification of a 634 kbp region, AR reconstructed a 430 kbp segmental tandem duplication supported by both AR and the AA breakpoint graph, containing oncogene GSE1 (Supplemental Fig.

S11a,b). This highlighted the ability of AR to also reconstruct classes of ultra-large, albeit less-complex SVs.

In the renal cancer cell line, CAKI-2, AA generated a breakpoint graph spanning 12.0 Mbp, joining regions from chr3 and chr12 (Supplemental Fig. S11c,d). Despite the lower overall copy number of this amplicon (~5), AR still reconstructed a 13.1 Mbp amplicon explaining 99.9% of the amplified genomic content in the AA-detected fCNA. Both amplicons for CAKI-2 and T47D appear to be intrachromosomal events given the AR results.

Across the focal amplifications we studied in seven cancer cell lines, we reported 64 individual amplified breakpoints detected by both AA and validated by AR (Supplemental Table S3). We also reported a summary of reconstruction findings for each sample and provided a list of reconstructed paths in Supplemental Table S4. Taken together, our data demonstrate the power of AR to combine NGS and OM data to elucidate a variety of complex fCNAs commonly found in cancer - enabling a deeper understanding of the fundamental mechanisms that give rise to fCNAs and promote cancer pathogenesis.

Discussion:

Revealing the architecture of fCNAs, particularly at a large scale, is critical to understanding their functional implications. For instance, rearrangements present in fCNAs can directly increase oncogene copy number, disrupt gene structure[41], and lead to dysregulation of chromatin[17–19]. Thus, understanding the organization and content of fCNAs is essential in predicting the behavior of the underlying sequences. Accurate reconstruction of fCNA architecture can provide insights into the mechanisms of their formation, leading to an improved understanding of the biological consequences of fCNA that would not be available solely from methods characterizing individual breakpoints.

While previous methods have characterized complex structural variation using both OM and NGS data[21,42], these methods have typically focused on the identification of individual variants and breakpoints[38]. AR represents a more robust and comprehensive algorithmic approach to reconstructing the fine architecture of a target fCNA. Indeed, while some of the individual junctions reported by AR in these cell lines were already known[21], by focusing on reconstructing entire amplicons through the propagation of breakpoint information into larger scaffolds, AR provides a deeper insight into the complex mechanisms that generate fCNA.

Genomic structural heterogeneity is problematic for any genome reconstruction, including focal amplifications and the structure of fCNA in stable cell lines may evolve over time. Despite the change in topology between linear HSR-like and circular ecDNA fCNAs, the breakpoint graphs between both circular and linear forms of the same samples are highly similar[6], suggesting ecDNA genomic structure is often not altered during reintegration. We further note that assembled OM contigs may fail to capture rare instances of structural heterogeneity in the genome. However, previous results suggest that focal amplifications conferring a fitness advantage to cancer cells are clonally amplified[5,43], allowing for an accurate reconstruction of the dominant structure.

AR produced a high-confidence reconstruction of the K562 BCR-ABL1 focal amplification, which is thought to be derived at least in part from a balanced translocation. Despite the presence of the AR-supported and FISH-validated HSR-like status of the BCR-ABL1 translocation in K562, there does not exist a completely validated model that explains the increased copy number of BCR-ABL1 in one single location. We cannot rule out the possibility that the BCR-ABL1 amplification in K562 is mediated through an ecDNA stage[44], given the transient nature of the emergence and retreat of ecDNA[15] and the highly rearranged genomic landscape surrounding BCR-ABL1.

The collection of paths reconstructed by AR represent possible reconstructions of the fCNA, and the collection of paths may contain multiple similar explanations for the fCNA architecture. This may be in part due to genomic heterogeneity, limitations of the optical map assembly process, or errors in linking scaffolds across overlapping graph segments. Furthermore, technological limitations related to the quality of OM assembly may affect the ability to reconstruct high-fidelity amplicons. Thus, identifying a single best path or collection of scaffolds which represent a reconstruction best explaining the breakpoint graph and OM data still requires some manual identification and interpretation. From the collection of output structures, AR does not automatically produce a prediction of the mechanism of amplification.

We have not yet adapted AR to accept breakpoint graphs generated by other tools or to accept breakpoint graphs derived from more balanced rearrangements - though the AR algorithm is designed to handle them if an accurate breakpoint graph was provided as input. Furthermore, recent advances in other long-range sequencing technologies[45] highlight the need to adapt the AR algorithm to work with more general long-read technologies – an aspect we plan to address in future development.

The accurate, multi-megabase scale, complex fCNAs reconstructed by AR not only describe fine structural features of fCNA architecture, but also reveal mechanistic signatures of fCNA formation, allowing for future interrogation of the relationship between fCNA architecture and the biological consequences of fCNA structure. In particular, methods to accurately characterize fCNAs will enable better classifications of cancer subtypes and their associated prognoses.

Methods

Cell culture

NCI-H460, K562, and HCC827 cells were obtained from ATCC and cultured in RPMI-1640 media supplemented with 10% FBS. HK301 cells were cultured as neural spheres in DMEM/F12 media supplemented with B27, EGF (20 ng/ml), FGF (20 ng/ml), and heparin (1 ug/ml). All cells were incubated under standard conditions.

Metaphase chromosome spreads

Metaphase cells were enriched by treating cells with Karyomax (Gibco) at a final concentration of 0.1µg ml^{-1}. Cells were collected, washed in PBS, and resuspended in 75mM KCl for approximately 15 minutes at 37°C. Cells were fixed by addition of an equal volume of Carnoy's fixative (3:1 methanol:glacial acetic acid). Cells were washed three additional times in Carnoy's fixative and dropped onto humidified glass slides.

FISH

Metaphase spreads were equilibrated in 2x SSC (30mM sodium citrate, 300mM NaCl, pH 7) for approximately 5 minutes. They were dehydrated using successive washes of 75%, 85%, and 100% ethanol for two minutes each and allowed to dry. FISH probes were diluted in hybridization buffer (Empire Genomics) and added to metaphase spreads on slides, along with 22mm^2 coverslips. Samples were denatured at 70-75°C for 30 seconds – 2 minutes. Probe hybridization was performed at 37°C for around 3 hours or overnight in a humid and dark chamber. Samples were washed successively in 0.4x SSC and 2x SSC with 0.1% Tween-20. Samples were incubated with DAPI (0.1µg ml^{-1} in 2x SSC) for 10 minutes, then washed with 2x SSC and briefly rinsed with H_2O. Samples were mounted with Prolong Gold, #1.5 coverslips, and sealed with nail polish.

Microscopy

Confocal microscopy was performed on a Leica SP8 Confocal microscope with white light laser and Lightning deconvolution. Fluorescent microscope images were acquired using an Olympus BX43 microscope with a QiClick cooled camera. Images were subsequently analyzed in ImageJ[46] (using the Bio-Formats plugin[47]), to perform cropping, add scale bars and perform global adjustments to image brightness.

Acquisition of WGS data

We previously published[5,6] WGS data to SRA for six of the seven cancer cell lines (GBM39, NCI-H460, HCC827, HK301, K562, T47D) analyzed here. For CAKI-2, we used WGS data published by the Cancer Cell Line Encyclopedia on SRA. A list of SRA accession numbers used is available in Supplemental Table 1.

Breakpoint graph generation

WGS data was aligned to hg19 with BWA-MEM[48] (version 0.7.17-r1188, default parameters) and the resulting alignments along with SNV calls produced by Freebayes[49] (version v1.3.1-17-gaa2ace8) were supplied as input to the Canvas[50] CNV caller (version 1.39.0.1598). The alignments and CNV seeds were filtered using AmpliconArchitect's amplified_intervals.py module. Seeds exceeding 40 kbp with copy number 5 were subsequently analyzed with AmpliconArchitect. AmpliconArchitect outputs a breakpoint graph encoding segmented CN calls and the discordant reads connecting the segments. We note that in most cases identical amplicon regions are identified when CNV caller ReadDepth[51] is used for seeding instead.

We standardized the breakpoint graph generation process into a workflow called PrepareAA, available on Github: https://github.com/jluebeck/PrepareAA. We used the default parameters specified by PrepareAA in this analysis. To produce *in silico* digestions of breakpoint graph segments into reference optical maps, we used the generate_cmap.py utility in AmpliconReconstructor. This method for *in silico* digestion can produce labeling patterns for the Bionano Saphyr DLE-1 labeling pattern, while many previous methods for *in silico* digestion do not.

OM data generation

High molecular weight (HMW) DNA was extracted from GBM39, HCC827, HK301, and K562 cells using the Bionano Prep Blood and Cell Culture DNA Isolation Kit (Bionano Genomics #80004), with minor modifications to recover good quality HMW gDNA. As detailed below, the Nick, Label, Repair, and Stain (NLRS) and Direct Label and Stain (DLS) reactions were carried out for the Bionano Irys and Saphyr platforms, respectively. To generate the Irys data, DNA was nicked using Nt.BspQI nicking endonuclease (NEB), followed by labeling, repairing, and staining, using the Bionano Prep NLRS DNA Labeling Kit (Bionano Genomics #80001) along with recommended NEB reagents. To generate the Saphyr data, DNA was labeled with DLE-1 enzyme, followed by proteinase digestion and a membrane clean-up step, using the Bionano Prep DLS DNA Labeling Kit (#80005). BspQI-labeled DNA was loaded onto the Irys Chip (Bionano Genomics #20249) and the run conditions were manually optimized on the Irys system (Bionano Genomics #30047) to ensure efficient DNA loading into the nanochannels. DLS-labeled DNA was loaded onto a Saphyr Chip (Bionano Genomics #20319), and run conditions were automatically optimized on the Saphyr system (Bionano Genomics #60239) using the Saphyr Instrument Control Software to maximize DNA loading. Raw images generated by Irys were processed into digital "Molecules" files using the Bionano software AutoDetect[25]. Images from the Saphyr system were processed into digital "Molecules" files via the Saphyr Instrument Control Software. For Irys data, molecules ≥150 kilobase pairs (kbp) were assembled into consensus genome maps using the Bionano Assembler[26,27] (version 5122), using default parameters; for Saphyr data, molecules ≥150 kbp were assembled into maps using Bionano Access (version 1.2.1)[26]. Bionano Genomics separately provided Saphyr OM data for cell lines K562, T47D, NCI-H460, and CAKI-2. The methods by which OM data was generated for those four cell lines were previously published[21].

Optical map contig alignments with SegAligner

SegAligner uses a dynamic programming (DP) approach to optical map alignment, with a recursion similar to previously proposed DP algorithms for OM alignment[52,53]. SegAligner scores OM alignments in a novel way which accounts for collapsed pairs of labels in the assembled OM contig and uses an E-value approach to compute alignment significance as method of controlling false alignments. We define label collapse as the phenomenon where two nearby labels on an OM contig or map are measured as a single label due to limitations of imaging[54].

SegAligner supports alignment of *in silico* digested segments of the reference genome (including entire chromosomes of the reference genome) and assembled optical map contigs. SegAligner supports models of error for data from both the Bionano Irys and Bionano Saphyr instruments, and we parameterize our methods for them separately (Supplemental Table 5). SegAligner also supports multiple modes of alignment including semi-global, fitting, and overlap alignment.

To motivate the notion of an OM alignment, we first define the concept of an OM matching region. Similarly to Valouev et al.,[53] a matching region is defined as the region between and including two labels on a map. For example, j and i in Supplemental Fig. S1b constitute a matching region with size $j - i$ and one unmatched label in-between. The alignment score for two matching regions depends on the size discrepancy of the matching regions and the number of unmatched labels in each matching region.

We define the following variables:

- *b* is a sorted list of real numbers corresponding to the positions of labels on the optical map contig in base pair units.
- *x* is a sorted list of real numbers corresponding to the positions of labels on a single *in silico* reference segment in base pair units.
- *P* is a matrix storing backtracking references
- *U* is a set storing reference segment label to contig label pairings which have already been used in previous iterations of the alignment process.
- *d* is the width of the band to consider for a banded alignment (default 6).
- *M* is a map which relates each label, *j* on a genomic segment, *x*, to the estimated probabilities for the left neighbor and right neighbor of *j,* that *j* and a neighbor would be observed as a single label (i.e. “collapse”).

Next, define $S[j][q]$ as the best score of aligning a subsequence of the first *j* labels on *b* with a subsequence of the first *q* labels on segment *x*, where *j* and *q* are included in the subsequences. Given two labels on the assembled contig *i, j*, and two labels on the reference genome segment *p, q* where *i* < *j,* and *p* < *q,* The DP recurrence used by SegAligner is (Algorithm 1)

$$S[j][q] = max_{\binom{max(0,j-d)\leq i<q}{max(0,q-d)\leq p<d}}\{S[i][p] + \ Score(i,j,p,q)\}$$

Where *Score* is the SegAligner scoring function for two OM matching regions. *Score* includes a function which computes the number of expected reference labels between p and q after accounting for label collapse. A backtracking matrix P is used to record the decision made in filling each cell S[j][q]. The DP Algorithm has complexity $O(mnd^2)$ where $m = |b|,\ n\ = |x|$ and d is the width of the band. Backtracking is performed in $O(m)$ steps by backtracking through the coordinates stored in *P*. We find a most-likely path by initializing the backtracking at $\operatorname{argmax}_{\mathrm{j,q}} S[j][q]$ or $S[|b|-1][|x|-1]$ for fitting alignment. Values used to parameterize the scoring function and label collapse map generation function given below are provided in Supplemental Table S1.

Algorithm 2: SegAligner scoring function

```
function Score(b,x,i,j,p,q,M):
    $f_n$ = c*(j - (i + 1))
    $e_{ref}$ = M(p,q)
    $f_p$ = c*$e_{ref}$
    $\Delta$ = $(abs((b[j] - b[i]) - (x[q] - x[p])))^k$
    return 2c - ($f_n$ + $f_p$ + $\Delta$)
```

As multiple regions of a long OM query might match similar regions of the reference, we extend the DP by masking out the best alignment path from the DP scoring matrix and recomputing the next best alignment.

Labels within approximately 2000 bp on an OM molecule may be read as a single label due to limitations of imaging, with increasing probability for smaller label-to-label intervals (Supplemental Fig. S1c). SegAligner captures that behavior in its scoring method, by precomputing the number of expected labels appearing in a collapsed label-set, given the reference.

To compute probabilities of label collapse, we assume a model in which the probability that a label at position *r* has merged with its right neighbor at position *s* is given by $P(r \to s) = \min\left(1, \left(\frac{(s-r)^t}{w^t}\right)\right)$. The map *M*, encoding the expected number of uncollapsed labels between two points on an *in silico* reference segment, is generated iteratively, by evaluating the following sum. $M(p,q)$ represents the sum of probabilities for each label between, but not including *p* and *q* that the label has collapsed with a neighbor. The sum of probabilities for [0,1] binary random variables to be 1 naturally gives the expected value of the sum of the binary random variables.

$$M(p,q) = \begin{cases} \sum_{p<k<q} \left(1 - \min\left(1, \frac{(x[k] - x[k-1])^t}{w^t}\right)\right)\left(1 - \min\left(1, \frac{(x[k+1] - x[k])^t}{w^t}\right)\right) & if\ x[q] - x[p] \geq \eta \\ 0\ if\ x[q] - x[p] < \eta \end{cases}$$

A genomic segment may appear multiple times in an optical map contig. Parameterizations of w, t and η are parameterized separately depending on the Bionano instrument used (Supplemental Table 5). SegAligner uses a set (*U*) to keep track of the pairings of segment labels (*q*) and reference labels (*j*) which form each significant high-scoring alignment. After a best-scoring alignment is found, the label pairings (*j,q*) are added to *U*. Subsequent alignments of that segment cannot re-use any pairings in *U.* This limits the creation of many nearly identical local alignments which differ by small indels, only one of which (the best scoring) is useful from a practical standpoint. We also placed a threshold on the number of times a single segment can be aligned to a single contig, so that low-complexity segments do not cause the aligner to stall (default 12).

Identifying significant high-scoring alignments

To compute statistically significant alignments, SegAligner uses a strategy similar to BLAST[55]. For each reference segment, *r*, SegAligner constructs a distribution of alignment scores representing the best scoring alignments of *r* to all contigs (Supplemental Fig. S1b). As this distribution may contain true alignments between *r* and one or more contigs, violating the random pairing assumption of the E-value model, SegAligner removes the highest 25 values from the distribution. From the remaining distribution of scores, we define a set of high scoring segment pairs (HSPs) which are the distribution of scores from the 85th percentile and up, from which SegAligner estimates parameters in the E-value model. We note that this region of the HSP scoring distribution tends to behave linearly (Supplemental Fig. S1c), allowing for a linear regression approach to parameter estimation.

SegAligner assigns an empirical E-value for each element in the sorted distribution of HSP alignment scores based on its rank (highest scoring having E-value 1). SegAligner then performs a local linear regression to estimate unknown variables in the E-value model. Generally, the E-value model is given by

$$E = Kmn_r e^{-\lambda S}$$

which implies

$$\log(E) = \log(Kmn_r) - \lambda S$$

where *m* is the size of the combined collection of contig labels, n_r is the number of labels on the reference segment, and *S* is the alignment score. As *K* and λ are unknown and represent the intercept and slope, respectively, SegAligner determines them from the empirical distribution of scores and E values using linear regression.

With all parameters known, the number of random high-scoring alignments, *a,* with score $\geq S$ is given by a Poisson distribution

$$P(a) - \frac{e^{-a}E^{a}}{a!}$$

This implies that finding at least one HSP for a given value of *E* is

$$P = 1 - e^{-E}$$

Thus, the score-cutoff S_r^* corresponding to a given probability, *P,* for segment *r*, is

$$S_r^* = \frac{-\log\left(-\frac{\log(1-P)}{Kmn_r}\right)}{\lambda}$$

SegAligner assigns to each reference segment a score which corresponds to the p-value cutoff for alignment significance. Default p-values are; 10^{-4} for semi-global alignment, 10^{-6} for overlapping alignment, and 10^{-9} for detection of new genomic reference segments aligning to contigs where the reference segment is not specified in the provided breakpoint graph segments (detection mode). SegAligner also computes the mean and median of segment-contig label pair alignment scores for each alignment exceeding the significance thresholds. Statistically significant scoring alignments failing mean and median thresholds (Supplemental Table 2) are filtered out. By default, AR attempts to align graph segments with at least 10 (Irys) or 12 (Saphyr) labels in the segment. However, the fitting mode of alignment only requires two endpoint labels, and so it is used in the path imputation step in AR.

AmpliconReconstructor – ARAlignDetect module

AmpliconReconstructor coordinates the alignment of in-silico digested breakpoint graph segments to optical map contigs using SegAligner (Fig. 1b). Alternately, AR can take as input XMAP-formatted alignments produced by other alignment tools. If OM contigs with alignments to graph segments contain unaligned regions with between 20 and 500 unmatched labels, and 200 kbp to 5 Mbp in length, those regions are extracted and searched against the reference genome. ARAlignDetect calls SegAligner in the "detection" mode, which then aligns the extracted unaligned region of the contig(s) to the specified reference genome. If significant alignments are found between unaligned regions of the contig and chromosomal segments in the reference, those segments are extracted, and their identity is added to the collection breakpoint graph segments. Finally, a new breakpoint graph is output containing the newly detected segments.

Reconstructing amplicon paths with AmpliconReconstructor

Optical map alignments of segments with contigs are converted into a scaffold, which we define as a collection of alignments where the genomic distance between each pair of alignment endpoints is known. AR represents the scaffolded alignments as a directed acyclic graph (DAG), where the nodes are an abstract representation of each OM alignment. Directed edges connect adjacent alignment endpoints. Overlapping alignments are connected with special directed edges referred to as "forbidden" edges (Fig. 1h). Two nodes are only connected by a non-forbidden edge if the right endpoint of the source node has one or fewer labels of overlap with the left endpoint of the destination node. Each contig with at least one alignment to a graph segment will comprise an individual scaffold.

Imputing paths in the scaffold with AmpliconReconstructor

Some segments in the breakpoint graph may be too short to be uniquely aligned to an OM contig. AR attempts to impute corrected paths in the scaffold using the structure of the breakpoint graph. For every non-forbidden edge in the scaffold graph with a gap size less than 400 kbp, AR identifies breakpoint graph nodes corresponding to the source and destination endpoints, which we will denote as *s,* and *t*. AR then uses a constrained depth-first search (DFS) strategy to identify paths in the breakpoint graph between *s* and *t*.

Finding all possible paths between two nodes may produce infinitely many solutions should a cycle exist between the two nodes, so the recursion is constrained to terminate if certain conditions are reached. The constraints used in the search procedure are:

1) The multiplicity of the segments in the candidate path must always remain less than or equal to the copy number of the segment as specified in the breakpoint graph.
2) If a candidate path reaches the destination vertex, its length in base-pair units must not be more than $\min(25000, 10000L_p)$ shorter than the distance between the source and destination vertices as expected given the scaffold backbone, where L_p is the length of the path in number of segments.
3) During path construction, the length of a candidate path must not exceed $\min(25000, 10000L_p)$ beyond the of the expected distance given the scaffold backbone.
4) The number of valid candidate paths connecting source to destination must not exceed 2^{10}.
5) The path may not form a trivial cycle from ultra-short breakpoint graph segments less than 100 bp long. Such cycles appearing in an NGS-derived breakpoint graph we assumed to be erroneous or artifactual.

As constraint #4 may cause failure of the DFS whereby a tractable number of paths is not found, AR implements a constrained BFS search as a fallback option, which is used when the DFS fails for that reason. By parsimony, shorter paths between two nodes are more likely to be correct, thus AR applies the same set of criteria for the BFS search, with the threshold in constraint #4 increased to 2^{16}.

All valid candidate imputation paths discovered by AR are scored by a fitting alignment procedure using SegAligner. To score a candidate path, the ordered path segments, as well as the first and last labels on the source and destination endpoints, are converted to a compound CMAP composed of the concatenated CMAPs of the individual segments. A fitting alignment is performed between the compound CMAP and the region of the contig between the alignment endpoints, using SegAligner. The path with the alignment score which most improves the junction score is kept. If no valid candidate path improves the score of the junction, it remains unimputed. The scaffold is then updated to contain the imputed breakpoint graph path.

Identifying linked scaffold paths with AR

Given the collection of scaffold DAGs, AR first searches for paths in the individual DAGs which represent “heaviest” paths in the scaffold DAG, where the weight of a path is the sum of the lengths of its segments in base pairs. AR stores the heaviest path(s) for each scaffold prior to performing scaffold linking.

AR leverages the two orthogonal sources of information encoded in the breakpoint graph and OM contigs to link individual scaffolds. As the breakpoint graph segments are not detected to contain interior breakpoints, two endpoint alignments of the same breakpoint graph segment may be linked across two contigs. AR searches for prefix paths and suffix paths in each DAG. From the collection of prefixes and suffixes, AR searches for overlap between scaffolds generated from different contigs. Given that a contig can be assembled in either direction, overlapping reverse oriented suffixes or prefixes can also be matched. AR exhaustively finds sub-paths hitting either end of a scaffold DAG, which have overlap with other endpoint sub-paths, where the endpoint sequence of the scaffold may be assembled in either direction.

Finding reconstructions in the linked scaffold graph

Given the graph of linked scaffolds, AR searches for paths in the graph which conform to the copy number ratios in the breakpoint graph. AR starts by searching for all paths in the graph which begin at endpoint nodes in the individual scaffolds. AR then uses a greedy approach to identify the longest unique paths which conform to the copy number restrictions. From the candidate paths, AR checks each path segment’s multiplicity against the copy numbers encoded in the breakpoint graph in a ratio-dependent manner.

AR iterates over all the segment multiplicities in the reconstructed path, and at each multiplicity level determines the maximum estimated genomic copy number of path segments with that multiplicity. If a path segment has a multiplicity that is greater than the genomic copy number of that segment divided by the maximum copy number of all segments with multiplicities less than the given segment, then the path violates the copy number ratio check. AR allows each segment in the reconstructed path to exceed by 1 copy the copy number expected given the ratio between breakpoint graph copy numbers and segment multiplicity. If n_p is the multiplicity of segment *n* in the candidate path, *P*, and n_g is the copy number of graph segment *n* in the breakpoint graph, then n_p must satisfy

$$n_p \leq \max\left(c, \frac{n_g}{m_g}\right) + 1, \qquad \forall\, n \in P$$

where

$$m_g = \max\left(i_g, \forall\, i \in P, i_p == c\right)$$

$$c \in \mathbb{Z}$$

$$n_p > c > 0$$

If a candidate path passes the copy number ratio check, it undergoes a pairwise comparison with other paths passing this criterion, to check for path uniqueness. A path is unique if it does not represent a subsequence of a previously identified unique path. Furthermore, no rotation of the path sequence may be a subsequence of a previously identified unique path. AR assess subsequence paths by computing a longest common substring between a candidate path and a previously identified unique path (Algorithm 3). As the paths are first sorted by total alignment score prior to the iterative approach, this method is a greedy algorithm which prioritizes long, heavy paths as being more likely to be identified as unique non-subsequence paths. AR categorizes paths as being cyclic if the first and last scaffold graph node in the path are the same, and the path length is greater than two, as this distinguishes cyclic paths from paths which appear cyclic such as singleton paths or paths which represent segmental tandem duplications. Paths reported by AR are output in the AmpliconArchitect “cycles” file format.

Algorithm 3: Greedy filtering of subsequence paths

```
Function FilterSubsequencePaths(sorted_paths):
  kept = empty array
  for P in sorted_paths do:
    isSubsequence = False
    for J in kept do:
      for R in the set of all rotations of path P:
        if R is a subsequence of J then:
          isSubsequence = True

    if not isSubsequence then:
      append P to kept
  return kept
```

Simulation of amplicons to measure AR performance

We used OMSim[56] (version 1.0) to simulate Bionano Irys OM data from the hg19 reference as well as from 85 non-trivial paths (i.e. not directly consistent with the reference genome) in AA-generated breakpoint graphs from 25 cancer samples, including both cyclic and non-cyclic breakpoint graph paths. OM molecules were simulated at 40x baseline coverage for each chromosome arm in hg19. The combined hg19 maps from all arms were assembled into a set of OM contigs using Bionano Assembler (version 5122). A similar process was performed using high-confidence breakpoint graph paths, which were converted to FASTA format and used for map simulation. For each simulated path, molecules were simulated at a range of copy numbers, and simulated molecules from the chromosome arm(s) (downsampled to the appropriate CN) from which the path segments came were combined and *de* novo assembled into OM contigs with BioNano Assembler. The resulting contigs from each amplicon simulation were combined with the previously simulated reference contigs and used as input to AR. For combination sets of three amplicons from the same sample, a similar downsampling and combination strategy was used, where molecules from each of the three amplicon simulations was separately downsampled based on the copy number settings of the mixture then combined. As heterogeneous combinations of amplicons may occur at different ratios, we selected three sets of copy numbers for this combination simulation cases: 20-20-20, 20-15-10, and 20-2-2.

Measuring AR simulation performance

We computed the longest common substring (LCS) between the AR paths and the ground-truth path and considered only the path having the LCS between AR and AA paths when computing precision and recall. We define the LCS here using the identities of the breakpoint graph segments and their orientations. We pre-filtered some possible assembly error reflected in the paths by removing ends of reconstructed paths which were trivial reconstructions of the reference genome and which were not supported by the AA path. To measure the accuracy of AR-reconstructed paths against the ground truth simulated paths, we developed a set of three measurements which were used in calculating performance and recall.

1) Length (bp): Reports the length of a breakpoint graph path in base pair units.
2) Nsegs: Reports the length of a breakpoint graph path in terms of the number of graph segments (unbiased towards genomic length)
3) Breakpoint: Reports the length of a breakpoint graph in terms of the number of breakpoint graph segment junctions in the path.

We define precision and recall as follows, where *M* is the path measurement function (Length (bp), Nsegs, or Breakpoint), *LCS* is the longest common substring function, P_{AA} is the sequence of segments in the AA path, and P_{AR} is the sequence of segments in the reconstructed AR path:

$$Precision\colon \frac{M\big(LCS(P_{AA}, P_{AR})\big)}{M(P_{AR})}$$

$$Recall\colon \frac{M\big(LCS(P_{AA}, P_{AR})\big)}{M(P_{AA})}$$

To summarize the precision and recall metrics in a single value, we computed a mean F1 score across all the simulated amplicons for a given set of simulation conditions as

$$mean\ F1 = \frac{\sum_i \left(2\,\frac{precision_i * recall_i}{precision_i + recall_i}\right)}{n}$$

Reconstructed path visualizations

We developed a visualization utility, CycleViz (https://github.com/jluebeck/CycleViz), which produces circular and linear visualizations of AR or AA reconstructed amplicons (Supplemental Fig. S2a,b), to create topologically correct visualizations of AR reconstructions. CycleViz accepts inputs including the path files reported by AR (in the AA “cycles” format) as well as the path OM alignment files (optional) and produces visualizations which show the reconstructed path, *in silico* digestion of the path segments and the alignments of the digested segments with assembled OM contigs. For circular and linear visualizations, CycleViz places path segments in the visualization based on the length of the segments and their position in the path. For circular visualization layouts, the relative positions are converted to polar coordinates and a circular layout is formed. We also developed a visualization utility for visualizing JSON-encoded scaffold graphs formed by AR using CytoscapeJS (Supplemental Fig. S2c).

Contributions:

J.L., V.B., and P.S.M conceived the work and designed the study. J.L. and V.B. developed the AmpliconReconstructor algorithm and software. C.C. and D.A.P. generated Bionano OM data and provided technical advice. J.L., S.R.D., and V.B. conceived and conducted the simulation study. J.T.L. and K.M.T. conducted FISH and microscopy experiments and provided technical advice. V.B., P.S.M., and J.A.L. supervised all experiments. V.D., C.Z., and U.R. performed computational analysis and provided technical advice. J.L., C.C., J.T.L., K.M.T., V.D., D.A.P., J.A.L., P.S.M., and V.B. wrote the paper.

Competing Interests:

P.S.M. and V.B. are co-founders of Boundless Bio, Inc. (BB), and serve as consultants. K.T. is currently employed by and receives income from BB. V.B. is a co-founder, and has equity interest in Digital Proteomics, LLC, and receives income from Digital Proteomics (DP). D.A.P. is employed by and receives income from Bionano Genomics, 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. BB and DP were not involved in the research presented here.

Acknowledgements:

The authors thank members of the Bafna and Mischel labs, as well as Dr. Marcy Erb (UCSD SOM Microscopy Core) for advice on FISH experiments, the Ecker Lab at the Salk Institute for use of the Bionano Irys optical mapping instrument, and Bionano Genomics, Inc., (Alex Hastie, Jian Wang, Ernest Lam, Andy Pang, and others) for supplying data and providing input on this project.

Code Accessibility:

The following tools are available online.

- AmpliconArchitect: https://github.com/virajbdeshpande/AmpliconArchitect
- PrepareAA: https://github.com/jluebeck/PrepareAA
- AmpliconReconstructor (& SegAligner): https://github.com/jluebeck/AmpliconReconstructor
- CycleViz: https://github.com/jluebeck/CycleViz
- ScaffoldGraphViewer: https://github.com/jluebeck/ScaffoldGraphViewer

References

1. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).

2. Bignell, G. R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898 (2010).

3. Stuart, D. & Sellers, W. R. Linking somatic genetic alterations in cancer to therapeutics. *Current Opinion in Cell Biology* **21**, 304–310 (2009).

4. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).

5. Turner, K. M. *et al.* Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* **543**, 122–125 (2017).

6. Deshpande, V. *et al.* Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat. Commun.* **10**, (2019).

7. Carroll, S. M. *et al.* Double minute chromosomes can be produced from precursors derived from a chromosomal deletion. *Mol. Cell. Biol.* **8**, 1525–1533 (1988).

8. Oobatake, Y. & Shimizu, N. Double-strand breakage in the extrachromosomal double minutes triggers their aggregation in the nucleus, micronucleation, and morphological transformation. *Genes, Chromosom. Cancer* gcc.22810 (2019). doi:10.1002/gcc.22810

9. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).

10. Jones, D. T. W. *et al.* Tandem duplication producing a novel oncogenic BRAF fusion gene defines the majority of pilocytic astrocytomas. *Cancer Res.* **68**, 8673–8677 (2008).

11. Menghi, F. *et al.* The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations. *Cancer Cell* **34**, 197-210.e5 (2018).

12. McClintock, B. The Stability of Broken Ends of Chromosomes in Zea Mays. *Genetics* **26**, 234–82 (1941).

13. Soler, D., Genescà, A., Arnedo, G., Egozcue, J. & Tusell, L. Telomere dysfunction

drives chromosomal instability in human mammary epithelial cells. *Genes Chromosom. Cancer* **44**, 339–350 (2005).

14. Kitada, K. & Yamasaki, T. The complicated copy number alterations in chromosome 7 of a lung cancer cell line is explained by a model based on repeated breakage-fusion-bridge cycles. *Cancer Genet. Cytogenet.* **185**, 11–9 (2008).

15. Nathanson, D. A. *et al.* Targeted Therapy Resistance Mediated by Dynamic Regulation of Extrachromosomal Mutant EGFR DNA. *Science (80-.).* **343**, 72–76 (2014).

16. Verhaak, R. G. W., Bafna, V. & Mischel, P. S. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nature Reviews Cancer* **19**, 283–288 (2019).

17. Wu, S. *et al.* Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature* (2019). doi:10.1038/s41586-019-1763-5

18. Morton, A. R. *et al.* Functional Enhancers Shape Extrachromosomal Oncogene Amplifications. *Cell* (2019). doi:10.1016/j.cell.2019.10.039

19. Mitsuda, S. H. & Shimizu, N. Epigenetic Repeat-Induced Gene Silencing in the Chromosomal and Extrachromosomal Contexts in Human Cells. *PLoS One* **11**, (2016).

20. Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* **20**, (2019).

21. Dixon, J. R. *et al.* Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50**, 1388–1398 (2018).

22. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol.* **15**, (2014).

23. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).

24. Dzamba, M. *et al.* Identification of complex genomic rearrangements in cancers using CouGaR. *Genome Res.* **27**, 107–117 (2017).

25. Cao, H. *et al.* Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *Gigascience* **3**, 34 (2014).

26. Software Downloads - Bionano Genomics. Available at: https://bionanogenomics.com/support/software-downloads/. (Accessed: 14th November 2019)

27. Anantharaman, T., Mishra, B. & Schwartz, D. Genomics via optical mapping. III: Contiging genomic DNA. *Proceedings. Int. Conf. Intell. Syst. Mol. Biol.* 18–27 (1999).

28. Leung, A. K.-Y. *et al.* OMBlast: alignment tool for optical mapping using a seed-and-extend approach. *Bioinformatics* btw620 (2016). doi:10.1093/bioinformatics/btw620

29. Anantharaman, T. S., Mishra, B. & Schwartz, D. C. Genomics via Optical Mapping II: Ordered Restriction Maps. *J. Comput. Biol.* **4**, 91–118 (1997).

30. Barr, L. F. *et al. c-Myc Suppresses the Tumorigenicity of Lung Cancer Cells and Down-Regulates Vascular Endothelial Growth Factor Expression 1*. *CANCER RESEARCH* **60**, (2000).

31. Cho, S. W. *et al.* Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. *Cell* **173**, 1398-1412.e22 (2018).

32. Vogt, N. *et al.* Amplicon rearrangements during the extrachromosomal and intrachromosomal amplification process in a glioma. *Nucleic Acids Res.* **42**, 13194–13205 (2014).

33. Storlazzi, C. T. *et al.* Gene amplification as doubleminutes or homogeneously staining regions in solid tumors: Origin and structure. *Genome Res.* **20**, 1198–1206 (2010).

34. Rowley, J. D. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243**, 290–293 (1973).

35. Virgili, A. & Nacheva, E. P. Genomic amplification of BCR/ABL1 and a region downstream of ABL1 in chronic myeloid leukaemia: A FISH mapping study of CML patients and cell lines. *Mol. Cytogenet.* **3**, (2010).

36. Chandran, R. K. *et al.* Genomic amplification of BCR-ABL1 fusion gene and its impact on the disease progression mechanism in patients with chronic myelogenous leukemia. *Gene* **686**, 85–91 (2019).

37. Grosveld, G. *et al.* The chronic myelocytic cell line K562 contains a breakpoint in bcr and produces a chimeric bcr/c-abl transcript. *Mol. Cell. Biol.* **6**, 607–616 (1986).

38. Zakov, S., Kinsella, M. & Bafna, V. An algorithmic approach for breakage-fusion-bridge detection in tumor genomes. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 5546–51 (2013).

39. Zakov, S. & Bafna, V. Reconstructing Breakage Fusion Bridge Architectures Using Noisy Copy Numbers. *J. Comput. Biol.* **22**, 577–594 (2015).

40. Kinsella, M. & Bafna, V. Combinatorics of the breakage-fusion-bridge mechanism. *J. Comput. Biol.* **19**, 662–678 (2012).

41. Koche, R. P. *et al.* Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. *Nat. Genet.* (2019). doi:10.1038/s41588-019-0547-z

42. Chan, E. K. F. *et al.* Optical mapping reveals a higher level of genomic architecture of chained fusions in cancer. *Genome Res.* **28**, 726–738 (2018).

43. Decarvalho, A. C. *et al.* Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. *Nat. Genet.* **50**, 708–717 (2018).

44. Morel, F. *et al.* Double minutes containing amplified bcr-abl fusion gene in a case of chronic myeloid leukemia treated by imatinib. *Eur. J. Haematol.* **70**, 235–9 (2003).

45. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).

46. Downloads - ImageJ. Available at: https://imagej.net/Downloads. (Accessed: 9th December 2019)

47. Linkert, M. *et al.* Metadata matters: Access to image data in the real world. *Journal of Cell Biology* **189**, 777–782 (2010).

48. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).

49. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. (2012).

50. Roller, E., Ivakhno, S., Lee, S., Royce, T. & Tanner, S. Canvas: Versatile and scalable detection of copy number variants. *Bioinformatics* **32**, 2375–2377 (2016).

51. Miller, C. A., Hampton, O., Coarfa, C. & Milosavljevic, A. ReadDepth: A parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One* **6**, (2011).

52. Huang, X. & Waterman ', M. S. Dynamic programming algorithms for restriction map comparison. **8**, 1–520 (1992).

53. Valouev, A. *et al.* Alignment of Optical Maps. *J. Comput. Biol.* **13**, 442–462 (2006).

54. Das, S. K. *et al.* Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Res.* **38**, (2010).

55. Karlin, S. & Altschul, S. F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 2264–8 (1990).

56. Miclotte, G. *et al.* OMSim: a simulator for optical map data. *Bioinformatics* **33**, 2740–2742 (2017).

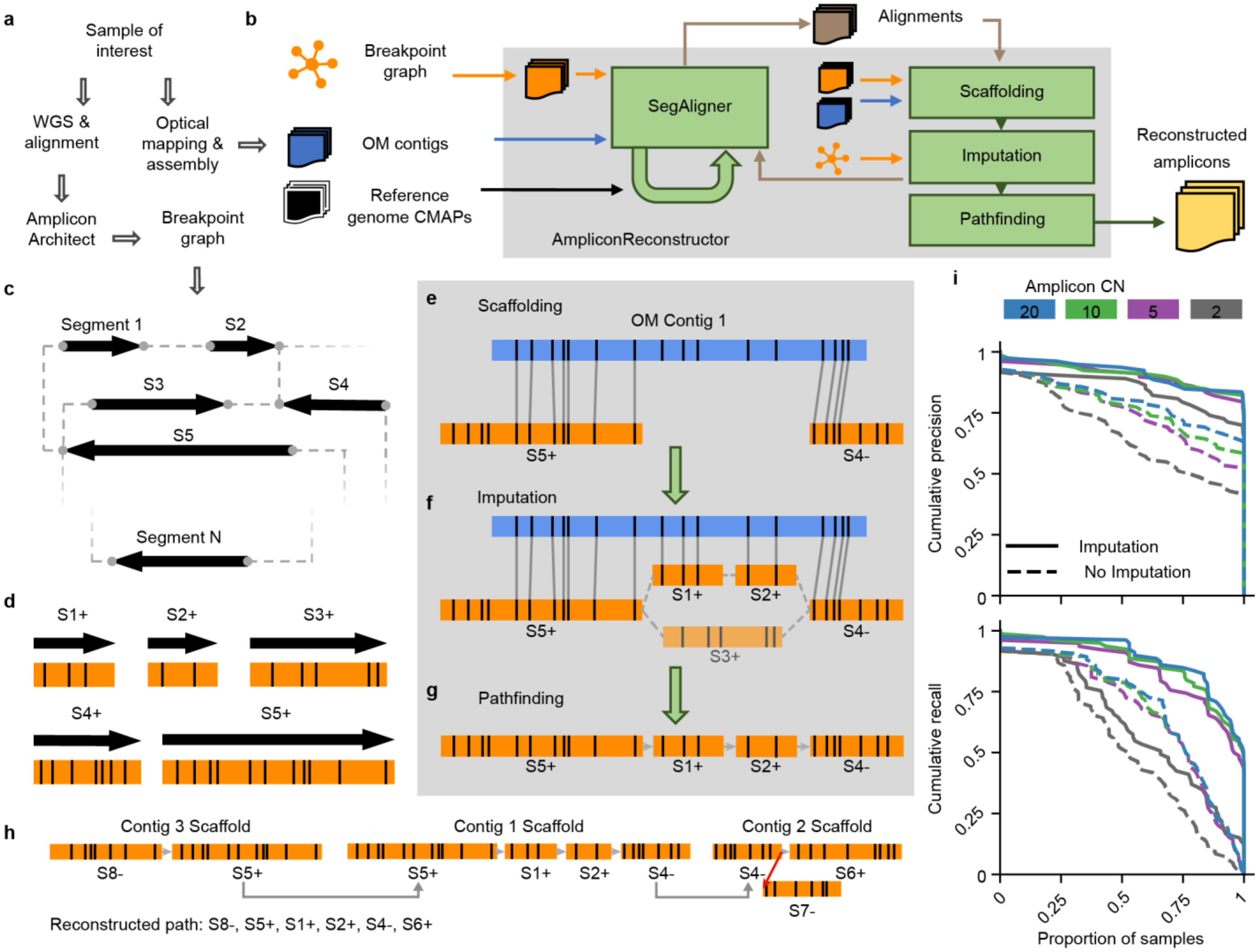

Figure 1: **[AmpliconReconstructor (AR) overview.] a**, Workflow to produce the necessary inputs for AR. AR accepts OM data in the consensus map (CMAP) format. **b**, High-level overview of the AR method, where the inputs and outputs are shown outside the grey box representing the AR wrapper. The green loop-back arrow on the SegAligner module represents the identification of reference segments not encoded in the breakpoint graph. **c**, A breakpoint graph with *N* segments. **d**, *In silico* digestion of breakpoint graph segments (orientation given by +/-) from **c** to produce graph OM segments. **e**, Alignment of graph OM segments to OM contigs produces a scaffold of segment-contig alignments. **f**, AR uses the structure of the breakpoint graph to identify paths between scaffold alignment endpoints which are also paths in the breakpoint graph. AR generates composite optical maps from combined path segments to score each candidate path against the gap in the scaffold. **g**, AR identifies a candidate path with maximum score out of the possible imputed paths between two alignments. **h**, AR links individual scaffolds sharing overlap between graph segments. The resulting graph has two types of edges, allowed (grey) and forbidden (red). **i**, Cumulative precision and recall curves based on simulated OM data for AR using SegAligner, calculated with the Length (bp) LCS metric. Line color indicates the copy number (CN) of the simulated amplicon.

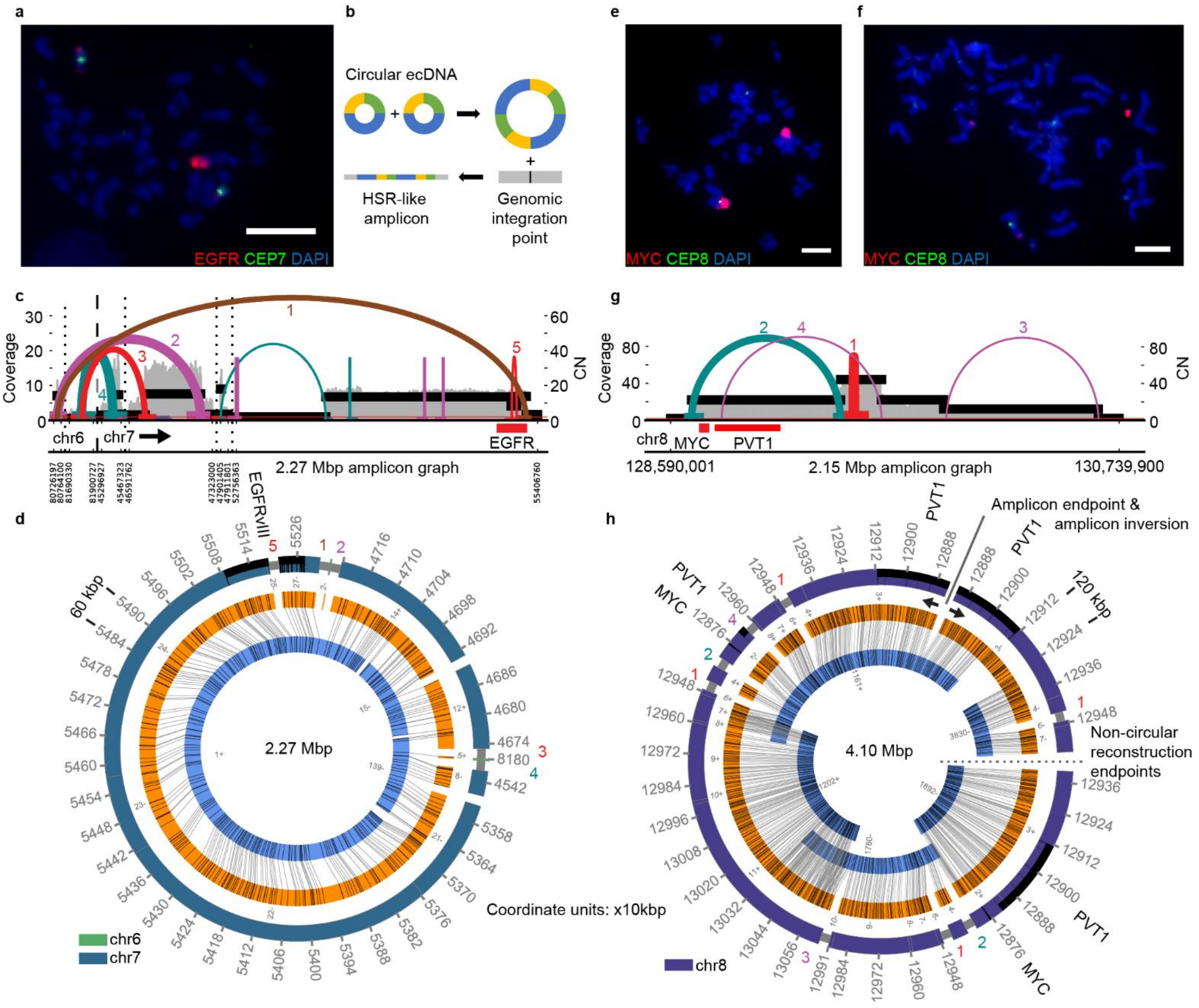

Figure 2: **[Reconstruction of extrachromosomal DNA (ecDNA)] a**, FISH with DAPI (4′,6-diamidino-2-phenylindole)-stained metaphase chromosomes in HK301 showing an HSR-like amplicon containing EGFR. Scale bar indicates 10 µm. **b**, Theoretical model for the integration of circular extrachromosomal DNA into HSR-like amplicons, preserving the structure of breakpoint graph. **c**, AA-generated breakpoint graph for HK301 containing EGFR and also segments from chr6. The coloring of the graph edges represents the orientation of the junction between the two segments. Edge thickness indicates AA-estimated breakpoint copy number. Vertical dashed lines separate segments from different chromosomes while dotted lines indicate distinct genomic regions from the same chromosome. Numbering of breakpoint edges corresponds with AR reconstruction breakpoint numbering. **d**, Cyclic AR reconstruction of HK301 amplicon containing EGFRvIII. Breakpoint graph edges supported by the AA graph are numbered in a manner corresponding to the numbering in panel **c**. **e**, FISH with DAPI-stained metaphase chromosomes in NCI-H460 shows HSR-like MYC amplicon. Scale bar indicates 7.3 µm. **f**, FISH with DAPI-stained metaphase chromosomes in NCI-H460 showing extrachromosomal MYC amplicon. Scale bar indicates 7.3 µm. **g**, AA-generated breakpoint graph for

NCI-H460 containing MYC and PVT1. **h**, AR reconstruction of the NCI-H460 amplicon. Indicated in this figure is an amplicon inversion point (top right) where the reconstruction explaining the full amplicon ends, and then the structure begins to repeat in the opposite direction (solid line & opposing black arrows). Also indicated is an endpoint for the non-circular reconstruction (center right) where the AR reconstruction and full amplicon structure both stop (dotted line).

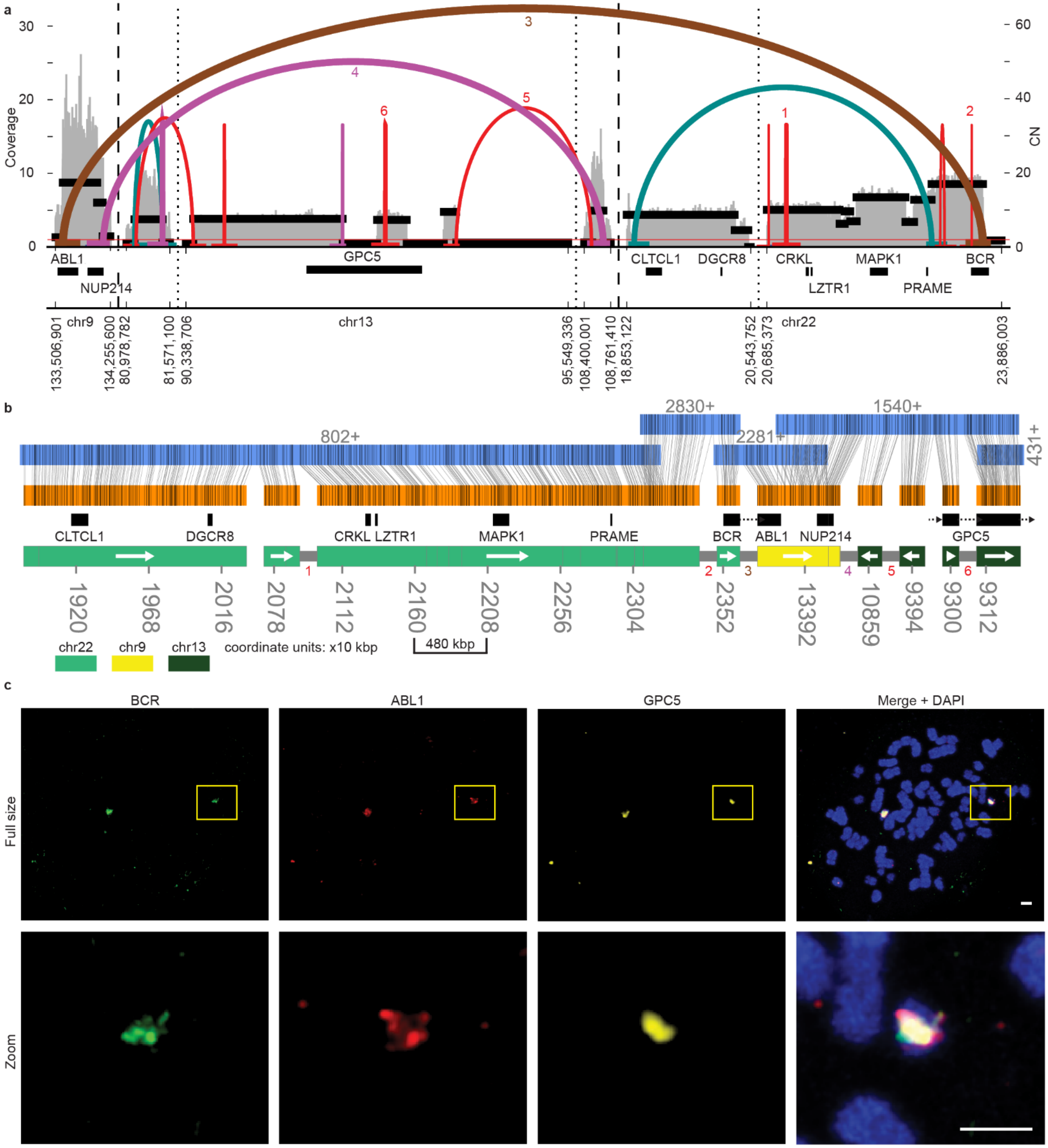
a
Coverage
0
10
20
30
CN
0
20
40
60
3
4
5
6
1
2
ABL1
NUP214
GPC5
CLTCL1
DGCR8
CRKL
LZTR1
MAPK1
PRAME
BCR
chr9
chr13
chr22
133,506,901
134,255,600
80,978,782
81,571,100
90,338,706
95,549,336
108,400,001
108,761,410
18,853,122
20,543,752
20,685,373
23,886,003
b
802+
2830+
2281+
1540+
431+
CLTCL1
DGCR8
CRKL LZTR1
MAPK1
PRAME
BCR
ABL1
NUP214
GPC5
1
2
3
4
5
6
1920
1968
2016
2078
2112
2160
2208
2256
2304
2352
13392
10859
9394
9300
9312
chr22
chr9
chr13
coordinate units: x10 kbp
480 kbp
c
BCR
ABL1
GPC5
Merge + DAPI
Full size
Zoom

Figure 3: [**Reconstruction of a complex Philadelphia chromosome] a**, AA-generated breakpoint graph for K562. Estimated copy number (CN), coverage, discordant reads forming breakpoint graph edges, and a subset of the genes in these regions are shown. **b**, AR reconstruction of an 8.5 Mbp focal amplification which was supported by both Irys and Saphyr reconstructions. The tracks from top to bottom are: OM contigs (with contig ID and direction indicated above), graph segments (alignments shown with vertical grey lines), gene subset and color-coded reference genome bar with genomic coordinates (scaled as 10 kbp units). Grey half-height bars between individual segments on the reference genome bar indicate support from edges in the AA breakpoint graph. White arrows inside the chromosome color bar indicate direction of genomic segment(s). Colored numbers correspond to numbered breakpoint graph edges in panel **a**. **c**, Multi-FISH using probes against BCR, ABL1 and GPC5 with DAPI-stained metaphase chromosomes. Scale bars indicate 2 µm in both “Full size” and “Zoom” rows.

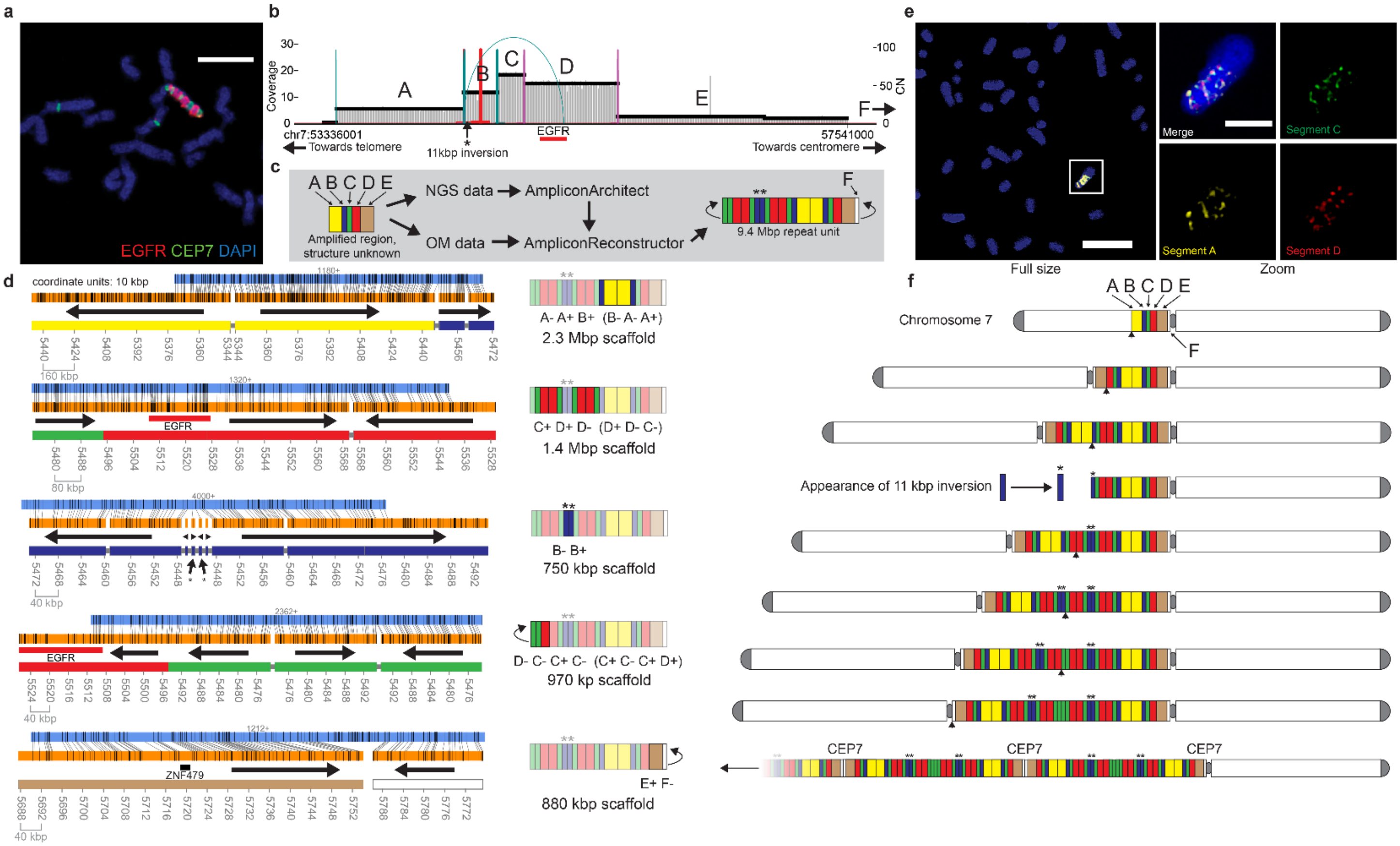

Figure 4: **[Reconstruction of Breakage-Fusion-Bridge.] a**, FISH confocal microscopy of DAPI-stained metaphase chromosomes in HCC827 showing multiple distinct bands of EGFR and CEP7 (chr7 centromeric repeat probe). Scale bar indicates 6µm. **b,** AA-generated breakpoint graph for amplified EGFR region in HCC827. Asterisk ('*') symbol indicates presence of 11 kbp inversion at 5' end of segment B. **c**, Workflow for analysis of amplified EGFR region in HCC827 to reveal BFB repeat unit structure. Amplified intervals detected by AA are labeled A-E and are colored yellow, blue, green, red and brown, respectively. "F" indicates a region identified by AR but not AA. **d**, Visualization of the AR-generated scaffolds (left column) and cartoon illustration of reconstructed region(s) of the BFB (right column), including segment sequence. Black arrows in the scaffold column indicate segment directionality. **e**, Multi-FISH for BFB segments using super-resolution confocal microscopy on DAPI-stained metaphase chromosomes in HCC827. FISH probes used for segments "A", "C", and "D" were RP11-64M3, RP11-117I14, and EGFR, respectively. Scale for full size image indicates 11 µm. Scale bar for zoomed images indicates 3 µm. Brightness was decreased using ImageJ between full size and zoomed images. **f**, Theoretical model of formation for HCC827 EGFR BFB. Each row indicates a prefix inversion and duplication characteristic of BFB, alongside other SVs. Black arrowheads beneath the intermediate step in each row indicates the breakpoint of the BFB chromosome. The bottom row shows multiple duplications of the BFB unit along with a pericentromeric region.