Descriptor

Patterns

FaNDOM: Fast nested distance-based seeding of optical maps

Highlights

- FaNDOM is a fast open-source aligner for OM data
- It utilizes a novel filtering strategy to reduce the search space of alignment
- The method enables discovery of large, complex genomic structural variants
- Structural variants suggested by FaNDOM include gene fusions and gene disruptions

Authors

Siavash Raeisi Dehkordi, Jens Luebeck, Vineet Bafna

Correspondence

vbafna@ucsd.edu (V.B.)

In brief

Optical mapping data is an orthogonal technique to DNA sequencing for the identification of genomic structural variants (SVs). We present a method, FaNDOM, which performs fast alignment of optical mapping data to the reference genome for identification of SVs. FaNDOM utilizes a novel filtering strategy, vastly reducing the search space of the alignment process, enabling rapid discovery of biologically interesting events.

Raeisi Dehkordi et al., 2021, Patterns 2, 100248
May 14, 2021
https://doi.org/10.1016/j.patter.2021.100248

CellPress

Patterns

CellPress
OPEN ACCESS

Descriptor

FaNDOM: Fast nested distance-based seeding of optical maps

Siavash Raeisi Dehkordi,[1,3,4] Jens Luebeck,[1,2,3] and Vineet Bafna[1,*]
[1]Department of Computer Science & Engineering, University of California, San Diego, La Jolla, CA 92093, USA
[2]Bioinformatics & Systems Biology Graduate Program, University of California, San Diego, La Jolla, CA 92093, USA
[3]These authors contributed equally
[4]Lead contact
*Correspondence: sraeisid@ucsd.edu (S.R.D.), vbafna@ucsd.edu (V.B.)
https://doi.org/10.1016/j.patter.2021.100248

THE BIGGER PICTURE Optical mapping (OM) is a rapidly maturing strategy for detecting large-scale rearrangements in genomes, leveraging ultra-long fragments of DNA imaged at very high depth of coverage (>100×). OM data reflect an orthogonal strategy to DNA sequencing, instead utilizing image-based detection of fluorescent tags associated with specific DNA motifs. The resulting data can be aligned back to the reference genome for discovery of genomic rearrangements and karyotypic abnormalities. Existing methods, however, are computationally demanding, making discovery harder. We present a novel method, FaNDOM, for alignment of OM data to the reference genome, and the additional discovery of structural variants. FaNDOM utilizes fast filtering algorithms based on constructing graph-based chains of seed matches, achieving orders of magnitude speedup, while maintaining high sensitivity, enabling a more comprehensive search of complex structural variations involving hundreds of kbp.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Optical mapping (OM) provides single-molecule readouts of fluorescently labeled sequence motifs on long fragments of DNA, resolved to nucleotide-level coordinates. With the advent of microfluidic technologies for analysis of DNA molecules, it is possible to inexpensively generate long OM data (>150 kbp) at high coverage. In addition to scaffolding for *de novo* assembly, OM data can be aligned to a reference genome for identification of genomic structural variants. We introduce FaNDOM (Fast Nested Distance Seeding of Optical Maps)—an optical map alignment tool that greatly reduces the search space of the alignment process. On four benchmark human datasets, FaNDOM was significantly (4–14×) faster than competing tools while maintaining comparable sensitivity and specificity. We used FaNDOM to map variants in three cancer cell lines and identified many biologically interesting structural variants, including deletions, duplications, gene fusions and gene-disrupting rearrangements. FaNDOM is publicly available at https://github.com/jluebeck/FaNDOM.

INTRODUCTION

Optical mapping (OM) is a rapidly maturing genome-mapping technology whose historical antecedents are at least a few decades old.[1] In the much older restriction-mapping technique, the use of sequence-specific restriction sites in a genome enabled unique "fingerprints" of the DNA. The initial restriction site maps were used to compare and position clones (genetic linkage maps) before sequencing.[2] Now, OM provides single-molecule readouts of the locations of fluorescently labeled sequence motifs on long fragments of DNA, resolved to nucleotide-level coordinates.[3] Despite the development of competing capillary sequencing and next-generation sequencing methods, optical maps continue to play an important role in scaffolding and assembly. With the advent of microfluidic technologies for high throughput of individual molecules and fluorescence-based visualization of covalently marked sites (labels), it is possible to generate high coverage (>100× of the human genome) with

long OM molecules (>150 kbp) for $500–1,000. For instance, the OM datasets analyzed in this paper had a median length of 191 kbp.

As the optical mapping technology evolves, the error profiles found in OM data also change. Bionano optical mapping (Bionano Genomics, San Diego, CA) uses direct covalent labeling of fluorescent molecules onto DNA fragments, as opposed to previous generations of OM, which used nickases. Its sources of error are orthogonal to DNA sequencing technologies,[4] and currently include incomplete labeling of donor sequences, false-positive labels, and imprecise resolution about exact locations of imaged labels. Other technology-specific phenomena, such as possible molecular chimerism or molecular stretching, also contribute to error. Computational methods, which handle OM data, must capture these various errors.

Given its uses for scaffold construction in *de novo* assembly projects,[5–7] optical mapping has matured to becoming a routine part of assembly pipelines for complex and/or large genomes. As a first step of this process, the OM fragments themselves are assembled into much larger (and error-corrected) OM contigs. The samples considered by our study had a median OM contig N50 of 38.4 Mbp. To achieve this, a computationally challenging problem of identifying overlapping OM fragments must be addressed. Much of the previous work about that problem uses dynamic programming algorithms to compare and align restriction maps,[8] and now extends to optical maps.[9,10] Newer methods, such as Kohdista[11] and MalignerIX,[12] tackle the overlapping fragment identification problems. Indexing and alignment-based methods have also been developed to map a sequence contig to a reference optical map genome, a requirement for scaffolding.[13,14]

Here, we consider the slightly different problem of mapping optical maps to a reference human genome for the purposes of identifying structural variants (SVs).[15,16] Such methods have been effective in identifying genomic abnormalities in Mendelian disease[17,18] as well as cancer.[19–21] Due to similar algorithmics, general methods for pairwise alignment or scaffolding, including Valouev,[22] SOMA,[23] TWIN,[13] and MalignerDP,[12] could be used in principle for mapping optical maps to an *in-silico*-digested reference genomic sequence. However, most of these methods do not repurpose well in practice, especially on data from the latest Bionano platform. Moreover, they do not call SVs. In contrast, OMBlast[24] and RefAligner[25] have previously demonstrated superior performance on Bionano data.[24,26] RefAligner specifically has been configured to call SVs. A new software, OMSV,[27] now combines RefAligner and OMBlast output to call SVs. Notably, RefAligner is a closed-source proprietary method, available only as pre-compiled binaries for specific hardware, and is very resource intensive, as described in the Results.

We introduce FaNDOM (Fast Nested Distance Seeding of Optical Maps)—an optical map alignment tool that introduces a novel method for seeding optical map alignments, greatly reducing the search space of the alignment process. FaNDOM is specifically optimized to handle data from the Bionano Saphyr optical mapping technology. The algorithmic and technology-specific improvements allow us to be significantly (4–14×) faster than competing tools while maintaining sensitivity and specificity. We used FaNDOM to map variants in three cancer cell lines and identified many structural variations, including deletion of tumor suppressor genes, duplications, gene fusions, and gene-disrupting rearrangements. FaNDOM is publicly available at https://github.com/jluebeck/FaNDOM.

RESULTS

As OMBlast[24,28] and RefAligner[25] were the best-performing pre-existing methods for mapping Bionano optical maps to a reference genome, we compared the performance of FaNDOM against Bionano RefAligner (Solve3.5.1) and OMBlast (OMTools v.1.4a). We also attempted to benchmark TWIN and Kohdista, but they are not specifically designed for this problem and did not perform as well (Methods S8).

Saphyr optical map data are publicly available for samples NA12878, GM09888, GM08331, and GM24143. We collected 270,000 raw molecules from each sample, where more than 85% of each molecule aligned to reference, as reported by Bionano, using their own RefAligner tool. We then ran FaNDOM, OMBlast, and RefAligner on this testing set.

Running time

We note that RefAligner is already highly optimized for the Saphyr technology, and is only provided as pre-compiled binary code, which runs on specific machine architectures. All experiments were conducted on an Intel(R) Core(TM) i9-9900 CPU @3.10GHz with 32 GB of main memory running Ubuntu 18.04.3 LTS (Bionic Beaver), using 10 threads. The results (Figure 1A) showed that FaNDOM was 4–6× faster than OMBlast and 13–14× faster than RefAligner on all datasets, highlighting the speedups created by our filtering methods. FaNDOM required approximately 2–2.5 GB of RAM for each thread (Methods S6). While OMBlast required less memory, the memory usage increased with increase in molecule size, and did not scale well for Saphyr-assembled contigs. The OMBlast documentation suggests 200 Gb RAM for mapping assembled contigs.

Mapping accuracy

We compared the accuracy FaNDOM, RefAligner, and OMBlast reported mappings on simulated and real data. Unlike DNA sequencing read mapping, which has discrete character matches and mismatches, it is not trivial to designate an OM molecule alignment as correct or incorrect on real data. Instead, we treated a mapping as correct if it was supported by at least two of the three methods.

We simulated datasets with "high" and "low" error (Methods S7), where high (H) corresponded to a false-positive label rate of 4 per 100 kbp, and stretch factor with standard deviation 0.02, which matched the Saphyr technology (Methods S7). Low (L) corresponded to a false-positive label rate of 1 per 100 kbp and stretch factor with standard deviation 0.01. All tools performed well on low-error. On high-error data, the three methods had very similar recall, with FaNDOM marginally higher, while FaNDOM precision lay between RefAligner and OMBlast (Figure 1B). On the cell lines, RefAligner had the highest precision and recall followed by FaNDOM and OMBlast. We note that RefAligner is better positioned to incorporate specifics of the Saphyr technology. The lower recall for FaNDOM relative to RefAligner can be partially attributed to the occasional removal of true maps during the filtering step. The precision can be

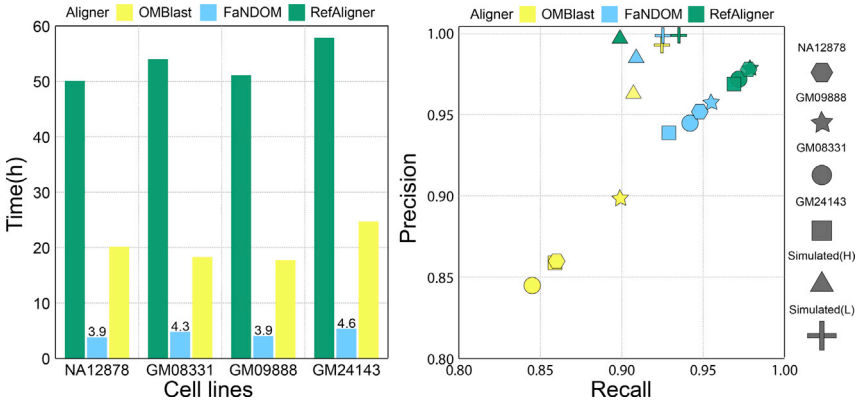


Figure 1. FaNDOM performance

(Left) Running time; (right) accuracy. The set of true positives (TP) were all mappings identified by at least two of the three methods. Recall = TP/(TP + FN), Precision = TP/(TP + FP).

improved by post-alignment filtering, and will be part of future release of FaNDOM after more datasets have been analyzed.

FaNDOM was 5× and 15× faster that OMBlast and RefAligner on cell lines (Figure 1A) as well as simulations (Figure S7; note different scale). As expected, simulations show that the running time increases with higher error rate for all methods.

SV detection

SV analysis continues to be a challenging problem requiring consensus from different methods and technologies. We compared the three methods using a benchmark of SV deletion calls of length >2,000bp on the genome NA12878. The benchmark was created previously using a multitude of technologies.[29] Figure 2A compares the performance of FaNDOM and RefAligner using assembled OM contigs. FaNDOM and RefAligner had comparable recall identifying 77% and 79% of the high-confidence calls, respectively, despite FaNDOM using filtering strategies to make the runtime faster by an order of magnitude. FaNDOM was much more aggressive in calling deletions compared with RefAligner. Spot checking, many of the FaNDOM-specific deletion calls appeared to be accurate (e.g., see Figure 2F).

OMSV[27] is another recent method for detecting SVs with OM data. It is an integrative tool that combines the output of RefAligner and OMBlast together, and is therefore even more compute intensive. As we could not run OMBLast on Saphyr contig data, we compared FaNDOM calls against pre-computed OMSV calls on NA12878 mapped to the hg38 reference and compared the calls with a benchmark deletion call set[15] on the hg38 reference (Figure 2B). The FaNDOM recall was 84% compared with the 70% recall of OMSV.

Detecting genomic insertions is one of the advantages of long-read technologies. FaNDOM predicted 719 insertions (Figure 2C). While there is no established call set of insertions for NA12878, 73% of the FaNDOM calls were previously reported as insertion polymorphisms in the Database of (human) Genomic Variants.[31] FaNDOM also identified a few ultra-long insertions in OM contigs (Figure 2D) that would be challenging with any competing technology due to the insertion size.

We investigated the FaNDOM-specific SV calls for possible error. The high-confidence dataset[29] has been collected by integrating a number of technologies, and is likely to be accurate. Nevertheless, many of its calls were discovered using short reads, while many of the FaNDOM-specific calls were >15 kbp (e.g., see Figure 2E). In addition, some of the FaNDOM-specific calls are in regions of low mappability (typically low complexity or repetitive sequence). Those breakpoints typically cannot be captured by short reads, but can be captured by long OM contigs (e.g., chr19:37,760K–37,795K; Figure 2F), demonstrating the complementarity of OM data to sequencing technologies. Moreover, assembled optical map contigs enable the detection of multiple breakpoints in one contig. As an example, Figure 2G represents an assembled OM contig from the K562 cell line that covers translocation from chr9 to chr13 and multiple breakpoints in chr13 spanning 500 kbp.

SVs in cancer cell lines

We ran FaNDOM on assembled OM contigs as well as OM molecules for cancer cell lines K562, CAKI-2, and H460—all of which are known to carry extensive rearrangements. Table1 summarizes some of the rearrangements identified by FaNDOM on assembled OM contigs. The rearrangements identified by FaNDOM, which included 1,800 large (>2 kbp) indels, 133 inter-chromosomal translocations, 28 fold-back reads, and 223 breakpoints that disrupted an existing gene, among other rearrangements. In this study, we focused specifically on genes that were deleted, and on translocations that disrupted or fused two genes.

The lung cancer cell line NCI-H460 has previously been documented to bear a focal amplification of the MYC/PVT1 region due to extrachromosomal DNA (ecDNA) and it has also been found to show evidence for intrachromosomal amplification in a homogeneously staining region (HSR).[32] Previous reconstruction of the MYC amplified region revealed a complex duplicated structure, which suggested that the ecDNA element containing MYC/PVT1 had reintegrated as an HSR in a non-native location.[20] The FaNDOM analysis identified a translocation from within the amplified ecDNA structure (chr8: 128,745, kbp) to a non-native location (chr12:7, 665k; Figure 3A) revealing chr12 to be the site of the HSR. Figure 3A also supports an inverted duplication at chromosome 8 as part of the amplified structure. In addition to recapitulating the breakpoints of the ecDNA, the FaNDOM analysis identified many partial or complete deletions of tumor suppressor genes, including LRP1B[33] (chr2: 141,735K–142,155K), TUSC7A[34] (non-coding; chr3: 116,295K–116,775K), FHIT[35] (chr3: 60,405K–60,735K), LSAMP[36] (chr3: 115,545K–116,145K). Notably, many of these deletions were on chr3. Many other rearrangements were identified providing a scenario of complex rearrangements in the cell line.

In the renal cancer cell line CAKI-2, we observed deletions or disruptions involving tumor suppressor genes, including CFHR1[37] (chr1: 196,665K–197,295K), RNF217 (chr6: 125,265K–125,505K),[38] RBFOX1 (chr16: 6,585K–7,155K),[39] FBXL7 (chr5: 15,825K–15,945K).[40] We also observed two fusions: TECRL1/GRIP1 (chr4: 65,205K, -, chr12: 66,975K, -) and RACGAP1/

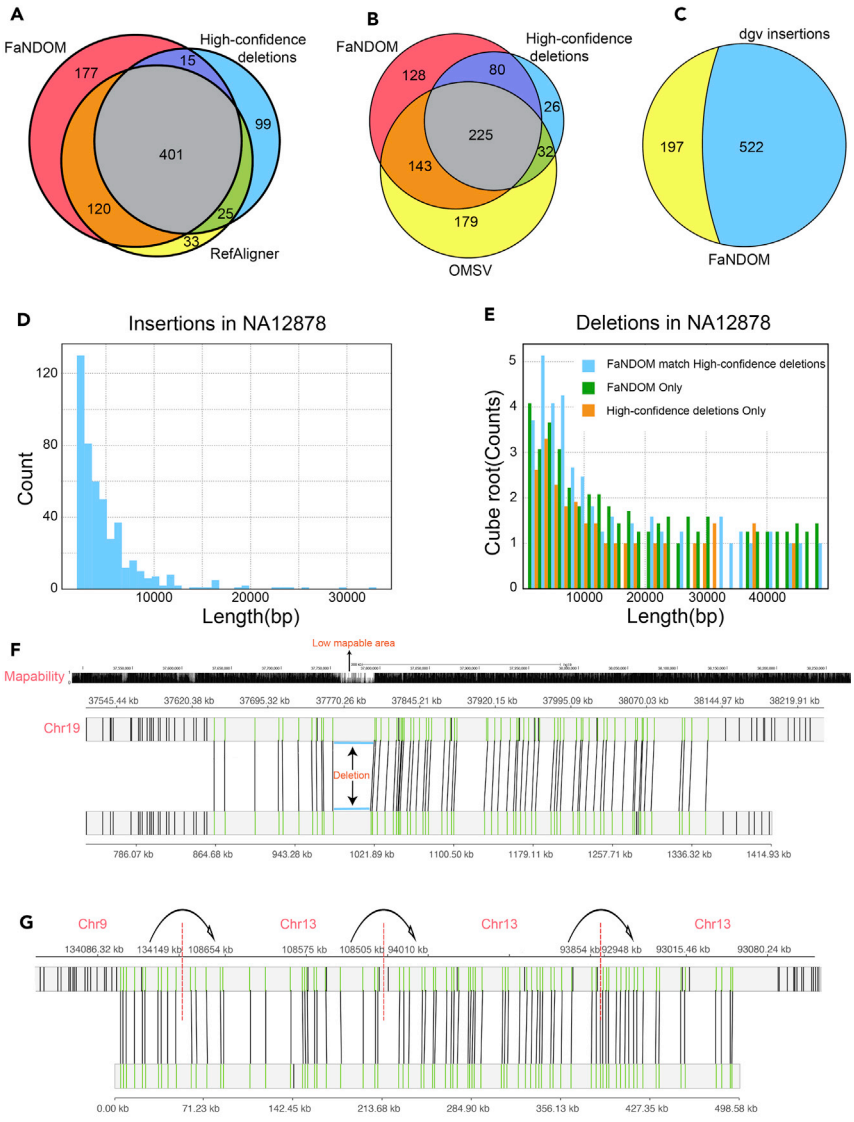


Figure 2. SV calling performance

(A) Comparison of FaNDOM and RefAligner deletion calls on NA12878 against a benchmark dataset from Parikh et al.,[29] using the hg19 reference.

(B) Comparison of FaNDOM and OMSV deletion calls on NA12878 against a benchmark created using multiple sequencing technologies published in Dixon et al.[15] using the hg38 reference.

(C) Insertions identified by FaNDOM for NA12878. The blue region signifies insertion polymorphisms identified by FaNDOM also in the Database of Genomic Variants.

(D) Length distribution of FaNDOM insertion calls for NA12878.

(E) Length distribution of FaNDOM and benchmark deletion calls (Parikh et al.[29]).

(F) A FaNDOM deletion not in the Parikh et al.[29] benchmark dataset likely due to its presence in a low mappability region.

(G) A FaNDOM alignment using assembled OM contigs that chains multiple breakpoints across 400 kbp on the K562 cell line. OM alignment visualizations were generated with MapOptics.[30]

AKAP6 (Figure 3B, chr12: 50,385K, -, chr14: 33,255K, +). RACGAP1 displays tumor malignancy potential[41] and is known to fuse with other genes, such as CERS5 and RAB34.[42]

K562 is a chronic myelogenous leukemia cell line with the Philadelphia chromosome. It was comprehensively analyzed recently using a multitude of technologies, including whole-genome sequencing and Hi-C.[43] FaNDOM confirmed some of the rearrangements of the previous study, such as the BCR-ABL1 fusion (Figure 3C), between chr22 and chr9. Among other rearrangements, we also observed an atypical microdeletion in 22q11, almost identical to a deletion previously associated with a congenital syndrome,[44] and a subset of a larger deletion reported for DiGeorge syndrome. The deletion encompasses the genes GSTT1, GSTT2, and GSTT2B, and deletions in these genes have previously been associated with esophageal cancer.[45]

While our results often matched the previously reported SVs,[43] there were a few notable differences. For example, in contrast with the previous finding of an inversion involving ORC6, MYLK3 on chr16, we observed a deletion (16:46, 725K–46,845K; Figure 3D) that partially removed ORC6 as well as a microinversion involving MYLK3. In a second example, the Zhou et al. study also identified a fusion of CDC25A/GRID1.[43] While we observe the same translocation, the directionality provided by the long reads suggests the disruption of the two genes, but not a fusion product (Figure 3E). We could confirm other chromosome 16 rearrangements, including an inverted duplication (88,605K–88,785 K), and another inverted duplication at chr13: 92,475K (Figure 3F).

DISCUSSION

Improvements to the optical mapping technology in terms of accuracy and cost has made it competitive for SV detection. At the same time, the raw data are harder to interpret and motivate the development of public domain tools for interpretation. In this paper, we focus on speeding up the mapping by relying on a novel filtering strategy that greatly improved speed without a significant loss of accuracy. The filtering relies on two ideas: (1) for most high-quality optical maps, it is relatively easy to find seeds that locate the reference target region for a query, and (2), by merging distances, thousands of queries can identify their target seeds in a single search-and-merge strategy. The results demonstrate the viability of this trade-off, leading to high speedup over other tools with only a small loss of sensitivity.

We recognize that our proposed method uses many parameters and, for the most part, the parameters are empirically determined to work for Saphyr. The optimal parameter values will be determined only after a large number of datasets have been analyzed, and will need to be retrained for newer technologies.

Table 1. Rearrangements in cancer cell lines

Cell lines	Indels	Interchromosomal translocations	Fold-back reads	Gene-disrupting breakpoints
CAKI-2	626	56	7	95
H-460	571	26	4	62
K562	603	21	17	66

In addition, non-human genomes, such as plants, may also require some significant recalibration of parameters and low-complexity annotations, which we have not yet explored. Nevertheless, because we have used FaNDOM to analyze many tens of thousands of molecules, the current choice of parameters appears to be robust for the current technology. Taken together, our results point to the value of using OM as a complementary technology for structural variation identification.

The detection of SVs is a key benefit of the OM technology, but it is harder to benchmark given the lack of large-scale, robust truth datasets. Our results suggest that FaNDOM can identify discordant alignments and breakpoints with high sensitivity. As many of the calls are based on cutoffs that can be adjusted, the results do not reveal any fundamental limitation of the filtering, but indicate a lack of additional calibration against a true gold standard. Additional analysis will be needed to identify systemic sources of false-positive calls.

We note that calling the structural variation mechanism itself is a secondary process that will require integration with other information, including copy-number changes, and this will be a topic of ongoing research. For example, one possible improvement includes pruning deletion calls by limiting results to the regions with a decrease in copy number consistent with heterozygous or homozygous deletion. With further improvements and methods development, OM technologies could be used to replace cytogenetics as a method of choice for revealing large-scale genetic abnormalities in Mendelian diseases and cancer.[17,20,21]

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Siavash Raeisi Dehkordi is the lead contact for this study and can be contacted by email at sraeisid@ucsd.edu.

Materials availability

This study did not generate any materials.

Data and code availability

The code for FaNDOM is available on GitHub at https://github.com/jluebeck/FaNDOM.

We used optical map data from the following individuals, and these data were obtained from the publicly available Bionano Saphyr datasets (https://bionanogenomics.com/library/datasets/)—NA12878, GM09888, GM08331, and GM24143. For cancer SV detection, we used previously published[20] Bionano Saphyr data from cancer cell lines K562, CAKI-2, and NCI-H460.

Method details

Conceptually, define an optical map as a sorted list of numeric values, representing the relative positions of labels on a fragment of DNA (Figure S1A). These numeric lists can be generated for any collection of individual OM molecules, assembled OM molecules, or from *in-silico*-predicted label positions on the reference genome. FaNDOM utilizes standard optical map data formats (.bnx or .cmap), where each imaged DNA fragment has been pre-converted to label position lists specified in base pair coordinates. An overview of the structure of the FaNDOM software is available in Figure S1B.

Pre-processing

Query fragments with length <25 kbp or containing less than 10 labels were filtered out from mapping. Similarly, queries containing consecutive labels with distance >250 kbp were removed (Methods S2).

Scaling refers to a systematic translation of physical inter-label distances into nucleotide distances. The Saphyr instrument performs a calibration to scale distances, estimating the number of base pairs present per image pixel. The process can on occasion be erroneous.[46] To recalibrate, FaNDOM randomly selects 250 molecules and estimates a corrected scaling factor using a grid search in a range of values between 0.96 and 1.2. The range was determined by experimenting from a set of 38 human samples (Methods S3). The rescaled molecules in each iteration are aligned to the reference. The scaling factor that achieves the highest total alignment score is selected for rescaling molecules before alignment.

Assembled OM contigs can be very large, often exceeding thousands of labels. As the alignment time grows quadratically with length, FaNDOM pre-processes assembled OM contigs by splitting them into smaller fragments, each containing 75 labels, with an overlap of 50 labels between endpoints of consecutive fragments. When alignment is completed, FaNDOM merges the alignments from overlapping fragments from assembled OM contigs to produce a complete alignment for the OM contig. In the case of conflicting alignments between overlapping contig fragments, FaNDOM maintains both partial alignments.

We convert the reference genome into a collection of expected label locations based on the *in silico* presence of the labeling motif throughout the reference. If the distance between two consecutive reference labels is less than 800 bp, they are replaced with the average of the two locations to account for the potential inability of resolving nearby OM labels (Methods S2). We also adapted a Bionano method[25] to identify and mask low-complexity regions in the human genome. Formally, denote a low-complexity region as containing at least five consecutive labels where the distance between adjacent labels is identical within 10% tolerance. Those could result in spurious alignments and are masked out. Specifically, in reference genome build hg19, 1.5 Mbp, which (0.04% of total reference genome) was masked out, while in hg38, 2.8 Mbp (0.09% of the total reference genome) was masked out (see Table S1 for masked regions).

Optical map alignment

The crux of a mapping procedure is an alignment of an optical map query to an *in silico* optical map of a reference sequence interval. The alignment maps query labels to the reference labels so that the inter-label distances between the query and reference are preserved (Figure S1).

The alignment of optical maps is a well-studied problem.[1,22] FaNDOM's scoring function follows previous methodologies, but diverges slightly. Consider reference R of length m and reference Q of length n labels. For $j \leq m$ and $q \leq n$, define $S[j][q]$ as the optimum score of aligning a subsequence (local alignment) ending at label j on R with a subsequence ending at label q on query Q. S can be computed using the following banded dynamic programming recurrence, where the band size is d:

$$S[j][q] = \max_{\substack{\max\{0,j-d\} \leq i<j, \\ \max\{0,q-d\} \leq p<q}} S[i][p] + \text{Score\_region}(R,i,j,Q,p,q), \qquad \text{(Equation 1)}$$

where, Score_region scores a match after penalizing for discrepancies in the match. Specifically, for $i<j, p<q$, let $f_n = (q - p - 1)$, $f_p = (j - i - 1)$ denote the number of unmatched labels in the query and reference, respectively. Then,

$$\text{Score\_region}(R,i,j,Q,p,q) = L - c(f_n + f_p) - |(R[j] - R[i]) - (Q[q] - Q[p])|^k.$$

We set $L = 10,000$ to represent a perfect match score. Empirical tests (Methods S4) indicated that a wide range of k, c showed identical performance. Increasing k, c resulted in the same alignments but with tighter boundaries. We chose the distance scale parameter $k = 1.15$ and false-label parameter $c = 3,000$ (Figure S4). After computing initial alignments for molecules, FaNDOM then identifies molecules, which are candidates for local/partial alignment discovery, as a prelude to SV analysis. In this partial

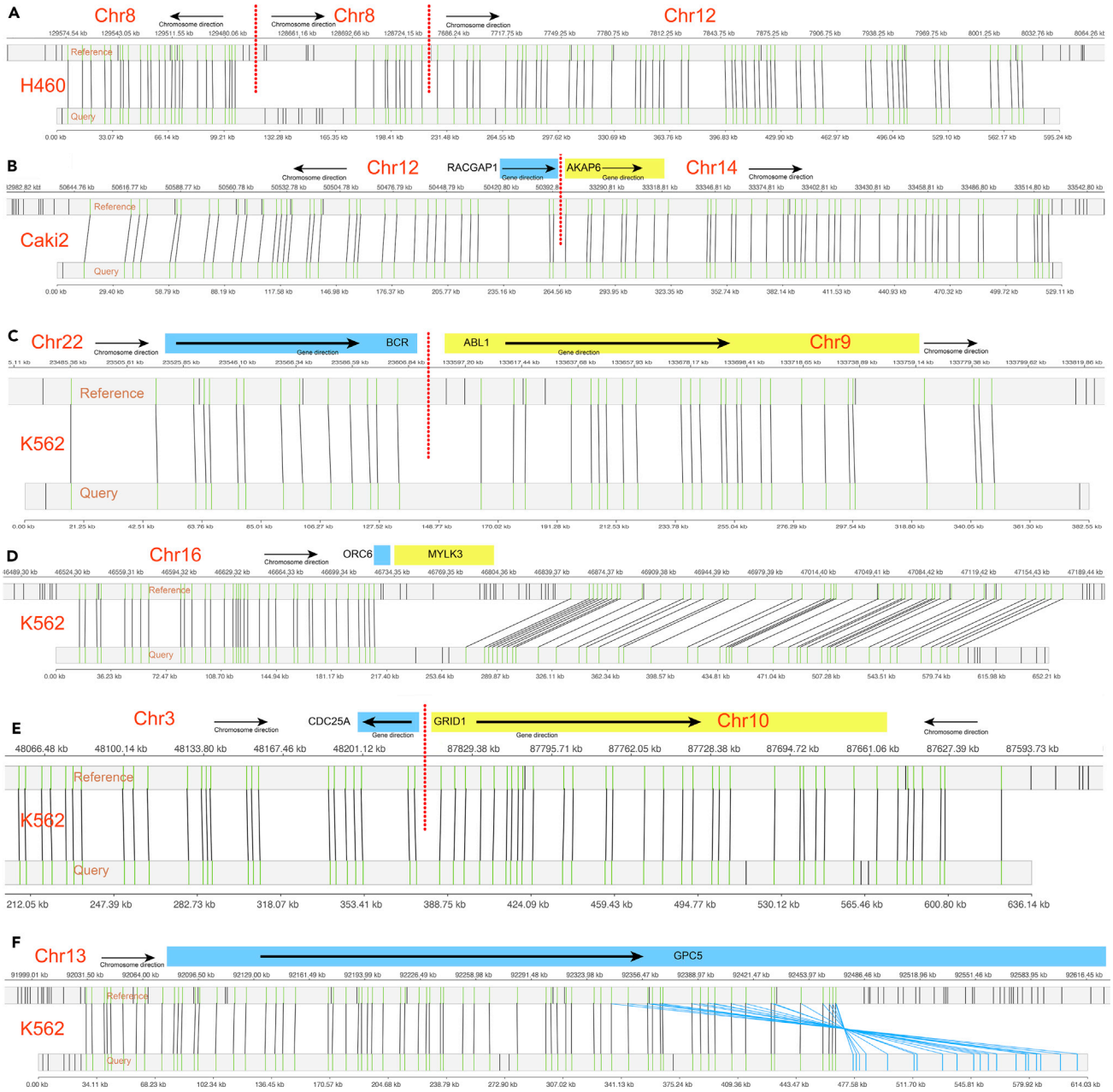


Figure 3. Examples of detected structural variants in cancer cell lines

(A) The chr8-chr12 translocation shows the integration of an Myc carrying ecDNA molecule onto chr12 in H460.

(B) A RACGAP1-AKAP6 fusion on CAKI-2.

(C) The BCR-ABL1 fusion on K562.

(D) Deletion of the genes ORC6 and MYLK3 with a partial inversion.

(E) A translocation that disrupts CDC25A and GRID1 but the direction is inconsistent with a fusion event.

(F) A "fold-back" inversion that duplicates and inverts GPC5 in K562.

alignment mode (see computing partial alignments for SV detection section below), where split-molecule alignments are allowed, FaNDOM computes more stringent partial alignments ($c = 7,500$, $k = 1.4$).

Alignment running time suggests the necessity of filtering.

The ungapped alignment algorithm has complexity $O(mnd^2)$. Despite algorithmic improvements and optimizations, our empirical results suggested that aligning a collection of two million OM fragments representing (100×) whole-genome coverage against every position on the human genome would take $\sim 700,000$ cpu-h. While assembly of OM fragments into contigs reduces the number of query sequences, the OM contigs are longer and the estimated time remains $\sim 15,000$ cpu-h. Therefore, similar to the Bionano RefAligner[25] and OMBlast,[24,28] we deploy a filtering strategy, where, for each query molecule, the goal is to identify a small collection of reference intervals to align the query with. The filter must be fast, sensitive (defined by the probability of the true reference location being included in the filtered reference intervals), and efficient (defined by the number of filtered regions per query—smaller being

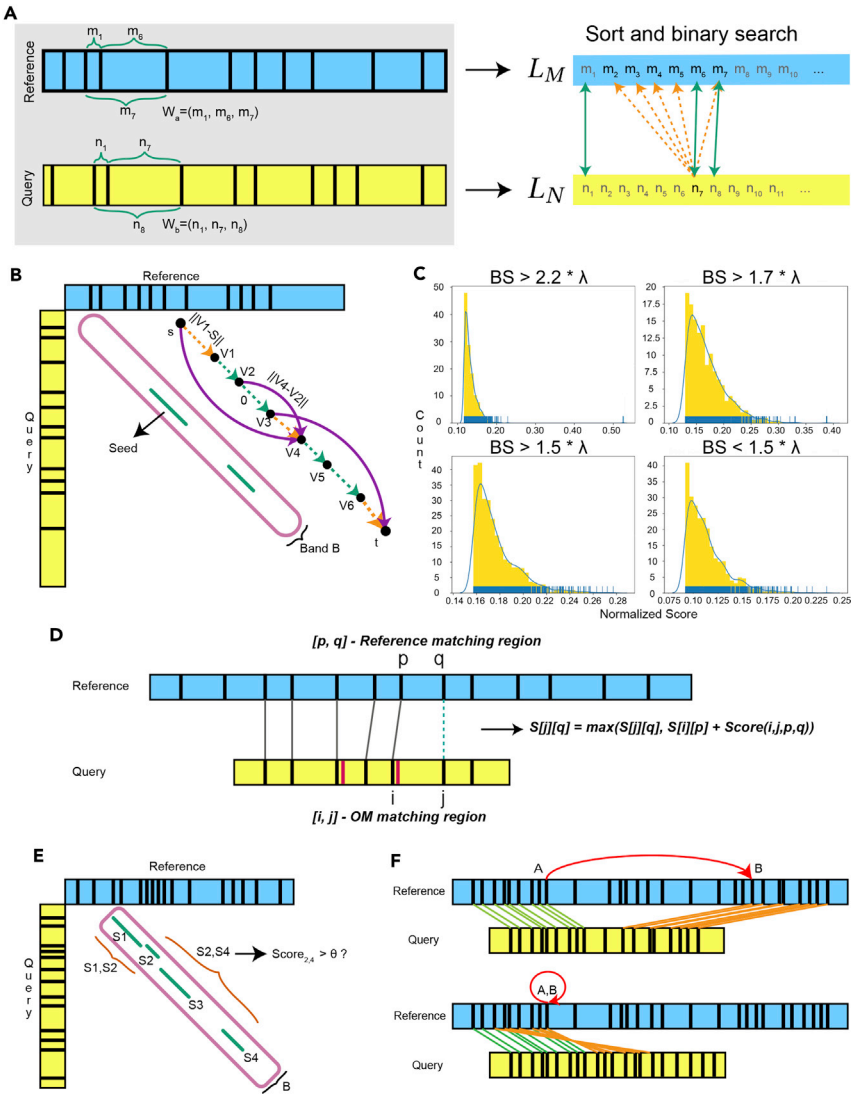


Figure 4. The FaNDOM workflow

(A) Search-and-merge filtering step in which genomic distances extracted from windows (W_a, W_b) and added to lists L_M and L_N. The lists L_M and L_N are merged and seeds are identified.

(B) Packing seeds into bands step, in which for each band B seeds inside it are formed into a directed acyclic graph, G, and the band is scored by finding the shortest path from s to t.

(C) Different score threshold possibilities for the band score distribution of bands for a single query. The best score is denoted as "BS."

(D) Dynamic programming for the alignment module in FaNDOM.

(E) Seed selection for partial alignment, which scores bands based on the shortest path between each pair of seeds inside the band B.

(F) SV detection module, which finds breakpoints based on multiple partial alignments. The alignment on top shows a breakpoint from A to B, the lower alignment visualizes an inversion, or "fold-back."

better). The filtered regions, or *seeds* are used to compute alignments and return the full or partial mappings of each query OM fragment or contig.

Search-and-merge filtering for optical maps

The key idea of filtering is that in a correct alignment there are some parts of query and reference, which are highly similar to each other, or that *all* inter-label distances in those regions are practically equivalent. Let $R[i,j]$ (respectively, $Q[i,j]$), denote the genomic distance between labels i,j in R (respectively, Q). Denote a window W_a in the reference as a collection of distances $R[i,j]$ for all $a \le i < j < a+3$. Windows W_b, in the query OMs are defined similarly. Let

$$W_b \xrightarrow[match]{} W_a \Leftrightarrow \forall_x \in W_b, \quad \exists y \in W_a : |x-y| \le T.$$

A default value of $T = 350$ was chosen empirically (Methods S4). In the search-and-merge procedure, we sort all genomic distances from every window of the reference (typically a chromosome) to a list L_m (Figure 4A). Similarly, for a collection of query OMs, we merge all sorted distances from all windows of each query in the collection into list L_n. Each distance $x \in L_m$ (respectively, $y \in L_n$) is associated with all reference windows (respectively, query windows) containing distance x (respectively, y).

Next, the sorted lists L_m, L_n are "search-merged" (Figure 4A). For each element $x \in L_n$ we perform two binary searches to identify the smallest and largest distances $y_1, y_2 \in L_m$ such that $x - y_1 \notin T, y_2 - x \notin T$. For all "matches" (x,y) where $y_1 \notin y \notin y_2$, we increment the match score of all window pairs associated with x and y. Finally, for all reference labels a, query labels b, such that $W_b \xrightarrow[match]{} W_a$, a *seed* (a, b, o), is generated, with $o \in \{+, -\}$ representing direction of match.

Packing seeds into bands

For each reference label a, and each query OM, FaNDOM explores a diagonal band B_a around a of width B_w (default value $B_w = 12,000$; Methods S4). Label a is filtered out if B_a contains fewer than $T_h = 4$ seeds (Methods S4). For retained bands, an edge-weighted directed acyclic graph G is constructed as follows: each node u in G corresponds to a pair of (query, reference) labels $\langle u_q, u_r \rangle$, where u_q (respectively, u_r) represents the nucleotide distance of the query label (respectively, reference label) from the first query (reference) label. Also, add nodes $s = \langle 0, 0 \rangle$ and $t = \langle l, l \rangle$ corresponding to the start and end of band B_a. For each seed u in the band, designate nodes u_1, u_2, u_3 corresponding to start, middle, and end of the seed. With few exceptions, we use Euclidean distances for edge weights so that $w(u,v) = ||v-u|| = \sqrt{(v_r - u_r)^2 + (v_q - u_q)^2}$. Specifically,

1. For each seed u, add edges (u_1, u_2) and (u_2, u_3) with weights 0 each; edge (s, u_1) with weight $\sqrt{2}u_{1q}$, and edge (u_3, t) with weight $\sqrt{2}(\ell - u_{3q})$.
2. For each pair of seeds u, v such that $u_{3q} \ngeq v_{1q}$ and $u_{3r} \ngeq v_{1r}$, add edge u_3, v_1 with weight $||v_1 - u_3||$.
3. For each pair of seeds u, v such that $u_{2q} = v_{1q}$ and $u_{2r} \ngeq v_{1r}$, add edge u_2, v_1 with weight $||v_1 - u_2||$.

We use dynamic programming to compute the weight w_{st} of the shortest (least-weight) path from s to t in G. The score of band B_a is given by

$$\text{score}(B_a) = 1 - \frac{w_{st}}{||t-s||}$$

A similar process is used for seeds in the reverse direction, with $s = \langle 0, l \rangle$, $t = \langle l, 0 \rangle$. For each query OM, we save the highest scoring 150 bands.

As a first idea, we could align the query map with the reference region for each of the 150 bands, and still achieve high speed and sensitivity. However, we observed that, in some cases, the top-scoring bands were significantly more likely to yield true alignments than other high-scoring bands, and that

the correct region was near the tail of the band score distribution and could be identified without aligning every candidate. We empirically fit the band scores to an exponential distribution with parameter λ and used the following empirical guidelines for scoring (Methods S4). For each query

$$\max_a \text{score}(B_a) \begin{cases} >2.2\lambda & \Rightarrow \text{Align top 10 bands} \\ >1.7\lambda & \Rightarrow \text{Align top 50 bands} \\ >1.5\lambda & \Rightarrow \text{Align top 100 bands} \\ \text{otherwise} & \Rightarrow \text{Align top 150 bands} \end{cases}$$

A band that is selected for alignment is converted to reference alignment boundaries by using the reference coordinate s_r of the source node s, and the query molecule Q of length $|Q|$. Specifically, for a padding factor p (default $p = 1000$), the region $s_r - p$ to $s_r + |Q| + p$ on the reference is used to align to the query molecule.

Computing partial alignments for SV detection

We identify SVs in two steps. First, queries that are either (1) unaligned, (2) have a mean alignment score less than 5,000/label, (3) the alignment does not cover 80% of the query length, or (4) has a total alignment length $\ngeq 25$ kbp, are targeted for partial alignments. The banding procedure is identical. For partial alignments, we compute local shortest paths between all pairs of seeds u, v as long as $||v_3 - u_1|| \ngeq 20$ kbp and the path contains at least four labels. If the corresponding band score

$$\left(1 - \frac{w_{u_1,v_3}}{||v_3 - u_1||}\right) \ngeq 0.4,$$

then the region gets a score of $(v_{3q} - u_{1q})\left(1 - \frac{w_{u1,v3}}{||v3-u1||}\right)$, and the top 300 candidate regions, each designated by a pair of nodes, are selected for alignment and re-ranking. A gapped alignment module is used and, if the score exceeds a threshold, the partial or gapped alignment is reported.

FaNDOM currently identifies *discordant alignments* (defined below) and *breakpoints*, which form the core of any SV discovery strategy, and defers the calling of actual SVs to a subsequent script that can be customized by the user. Recall that an alignment is a chain of matches $(q_0,r_0),(q_1,r_1),\ldots,(q_t,r_t)$. For alignments below a threshold score, if there exists $0 \ngeq i<t$ such that (1) $|(q_{i+1} - q_i) - (r_{i+1} - r_i)| > 2000$, (2) $|(q_{i+1} - q_0)| \geq \max\{10000, 0.25 \cdot \text{querylength}\}$, and (3) $|(q_t - q_{i+1})| \geq \max\{10000, 0.25 \cdot \text{querylength}\}$, then a discordant alignment is called. Discordant alignments typically represent insertions/deletions, but may also represent small inversions flanked by high-quality alignments on both sides.

Breakpoints refer to a pair of coordinates that are non-adjacent on the reference, but are together on the query. Consider two partial alignments that involve the same query molecule, described by $A1 : (q_0, r_0), \ldots, (q_i, r_i)$ and $A2 : (q_j, r_j), \ldots, (q_t, r_t)$. Note that $r_0, \ldots, r_i$ could potentially be on a different chromosome than $r_j, \ldots r_t$. Define $o_i, o_j \in \{+, -\}$ using $o_i = \text{sgn}(r_i - r_0)$, and $o_j = \text{sgn}(r_t - r_j)$. FaNDOM calls a *breakpoint* (r_i, o_i, r_j, o_j) if there is no partial alignment involving the labels between q_i and q_j. Breakpoints are clustered if their endpoints are within 30 kbp, and each breakpoint is listed along with its "support," or the number of alignments consistent with the breakpoint. Subsequent scripts are used to describe the rearrangement that creates the breakpoint. For example, $(r_i, +, r_j, +)$ describes a homozygous (respectively, heterozygous) deletion if r_i and r_j are on the same chromosome and the fragment coverage in the interval $[r_i, r_j]$ is 0 (respectively, half of normal coverage).

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100248.

ACKNOWLEDGMENTS

The research was supported by a grant from the NIH (GM114362). We would like to thank Andy Pang of Bionano Genomics, Inc. for his feedback and assistance with data interpretation and explanations of the Bionano pipelines.

AUTHOR CONTRIBUTIONS

S.R.D., J.L., and V.B. designed the study, developed the algorithms, conducted analysis, and wrote the paper. S.R.D. and J.L. developed the code for FaNDOM.

DECLARATION OF INTERESTS

V.B. is a co-founder, consultant, and SAB member of and has equity interest in Boundless Bio, Inc. (BB) and Digital Proteomics, LLC (DP) and also receives income from DP. 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.

Received: January 20, 2021
Revised: March 8, 2021
Accepted: April 1, 2021
Published: May 3, 2020

REFERENCES

1. Schwartz, D.C., Li, X., Hernandez, L.I., Ramnarain, S.P., Huff, E.J., and Wang, Y.K. (1993). Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. Science *262*, 110–114.
2. Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. *32*, 314–331.
3. Lam, E.T., Hastie, A., Lin, C., Ehrlich, D., Das, S.K., Austin, M.D., Deshpande, P., Cao, H., Nagarajan, N., Xiao, M., and Kwok, P.Y. (2012). Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat. Biotechnol. *30*, 771–776.
4. Chen, P., Jing, X., Ren, J., Cao, H., Hao, P., and Li, X. (2018). Modelling BioNano optical data and simulation study of genome map assembly. Bioinformatics *34*, 3966–3974.
5. Zhou, S., Wei, F., Nguyen, J., Bechner, M., Potamousis, K., Goldstein, S., Pape, L., Mehan, M.R., Churas, C., Pasternak, S., et al. (2009). A single molecule scaffold for the maize genome. PLoS Genet. *5*, e1000711.
6. Teague, B., Waterman, M.S., Goldstein, S., Potamousis, K., Zhou, S., Reslewic, S., Sarkar, D., Valouev, A., Churas, C., Kidd, J.M., et al. (2010). High-resolution human genome structure by single-molecule analysis. Proc. Natl. Acad. Sci. U S A *107*, 10848–10853.
7. Pan, W., Jiang, T., and Lonardi, S. (2020). OMGS: optical map-based genome scaffolding. J. Comput. Biol. *27*, 519–533.
8. Huang, X., Waterman, M.S., and Oct. (1992). Dynamic programming algorithms for restriction map comparison. Comput. Appl. Biosci. *8*, 511–520.
9. Anantharaman, T.S., Mishra, B., and Schwartz, D.C. (1997). Genomics via optical mapping. II: ordered restriction maps. J. Comput. Biol. *4*, 91–118.
10. Valouev, A., Li, L., Liu, Y.-C., Schwartz, D.C., Yang, Y., Zhang, Y., and Waterman, M.S. (2005). Alignment of optical maps. In Research in Computational Molecular Biology, S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P.A. Pevzner, and M. Waterman, eds. (Springer Berlin Heidelberg), pp. 489–504.
11. Muggli, M.D., Puglisi, S.J., and Boucher, C. (2019). Kohdista: an efficient method to index and query possible Rmap alignments. Algorithms Mol. Biol. *14*, 25.
12. Mendelowitz, L.M., Schwartz, D.C., and Pop, M. (2016). Maligner: a fast ordered restriction map aligner. Bioinformatics *32*, 1016–1022.
13. Muggli, M., Puglisi, S.J., Boucher, C., 2014. Efficient indexed alignment of contigs to optical maps, 68–81.
14. Leinonen, M., and Salmela, L. (2020). Optical map guided genome assembly. BMC Bioinformatics *21*, 285.
15. Dixon, J.R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V.T., Yardimci, G.G., Chakraborty, A., Bann, D.V., Wang, Y., et al. (2018). Integrative detection and analysis of structural variation in cancer genomes. Nat. Genet. *50*, 1388–1398.
16. Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat. Commun. *10*, 1784.
17. Barseghyan, H., Tang, W., Wang, R.T., Almalvez, M., Segura, E., Bramble, M.S., Lipson, A., Douine, E.D., Lee, H., Delot, E.C., et al. (2017). Next-

generation mapping: a novel approach for detection of pathogenic structural variants with a potential utility in clinical diagnosis. Genome Med. *9*, 90.

18. Dai, Y., Li, P., Wang, Z., Liang, F., Yang, F., Fang, L., Huang, Y., Huang, S., Zhou, J., Wang, D., et al. (2020). Single-molecule optical mapping enables quantitative measurement of D4Z4 repeats in facioscapulohumeral muscular dystrophy (FSHD). J. Med. Genet. *57*, 109–120.
19. Chan, E.K.F., Cameron, D.L., Petersen, D.C., Lyons, R.J., Baldi, B.F., Papenfuss, A.T., Thomas, D.M., and Hayes, V.M. (2018). Optical mapping reveals a higher level of genomic architecture of chained fusions in cancer. Genome Res. *28*, 726–738.
20. Luebeck, J., Coruh, C., Dehkordi, S.R., Lange, J.T., Turner, K.M., Deshpande, V., Pai, D.A., Zhang, C., Rajkumar, U., Law, J.A., et al. (2020). AmpliconReconstructor integrates NGS and optical mapping to resolve the complex structures of focal amplifications. Nat. Commun. *11*, 4374.
21. Neveling, K., Mantere, T., Vermeulen, S., Oorsprong, M., van Beek, R., Kater-Baats, E., Pauper, M., van der Zande, G., Smeets, D., Weghuis, D.O., et al. (2020). Next generation cytogenetics: comprehensive assessment of 48 leukemia genomes by genome imaging. bioRxiv. 2020.02.06.935742. https://doi.org/10.1101/2020.02.06.935742.
22. Valouev, A., Schwartz, D.C., Zhou, S., and Waterman, M.S. (2006). An algorithm for assembly of ordered restriction maps from single DNA molecules. Proc. Natl. Acad. Sci. U S A *103*, 15770–15775.
23. Nagarajan, N., Read, T.D., and Pop, M. (2008). Scaffolding and validation of bacterial genome assemblies using optical restriction maps. Bioinformatics *24*, 1229–1235.
24. Leung, A.K.-Y., Kwok, T.-P., Wan, R., Xiao, M., Kwok, P.-Y., Yip, K.Y., and Chan, T.-F. (2016). OMBlast: alignment tool for optical mapping using a seed-and-extend approach. Bioinformatics *33*, 311–319, https://doi.org/10.1093/bioinformatics/btw620.
25. Shelton, J.M., Coleman, M.C., Herndon, N., Lu, N., Lam, E.T., Anantharaman, T., Sheth, P., and Brown, S.J. (2015). Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. BMC Genomics *16*, 734.
26. Yuan, Y., Chung, C.Y., and Chan, T.F. (2020). Advances in optical mapping for genomic research. Comput. Struct. Biotechnol. J. *18*, 2051–2062.
27. Li, L., Leung, A.K., Kwok, T.P., Lai, Y.Y.Y., Pang, I.K., Chung, G.T., Mak, A.C.Y., Poon, A., Chu, C., Li, M., et al. (2017). OMSV enables accurate and comprehensive identification of large structural variations from nanochannel-based single-molecule optical maps. Genome Biol. *18*, 230.
28. Leung, A.K.-Y., Jin, N., Yip, K.Y., and Chan, T.-F. (2017). OMTools: a software package for visualizing and processing optical mapping data. Bioinformatics *33*, 2933–2935, https://doi.org/10.1093/bioinformatics/btx317.
29. Parikh, H., Mohiyuddin, M., Lam, H.Y., Iyer, H., Chen, D., Pratt, M., Bartha, G., Spies, N., Losert, W., Zook, J.M., et al. (2016). svclassify: a method to establish benchmark structural variant calls. BMC Genomics *17*, 64.
30. Burgin, J., Molitor, C., and Mohareb, F. (2019). MapOptics: a light-weight, cross-platform visualization tool for optical mapping alignment. Bioinformatics *35*, 2671–2673.
31. MacDonald, J.R., Ziman, R., Yuen, R.K., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. *42*, D986–D992.
32. Turner, K.M., Deshpande, V., Beyter, D., Koga, T., Rusert, J., Lee, C., Li, B., Arden, K., Ren, B., Nathanson, D.A., et al. (2017). Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature *543*, 122–125.
33. Liu, C.X., Li, Y., Obermoeller-McCormick, L.M., Schwartz, A.L., and Bu, G. (2001). The putative tumor suppressor LRP1B, a novel member of the low density lipoprotein (LDL) receptor family, exhibits both overlapping and distinct properties with the LDL receptor-related protein. J. Biol. Chem. *276*, 28889–28896.
34. Li, N., Shi, K., and Li, W. (2018). TUSC7: a novel tumor suppressor long non-coding RNA in human cancers. J. Cell Physiol. *233*, 6401–6407.
35. Waters, C.E., Saldivar, J.C., Hosseini, S.A., and Huebner, K. (2014). The FHIT gene product: tumor suppressor and genome "caretaker". Cell Mol. Life Sci. *71*, 4577–4587.
36. Kresse, S.H., Ohnstad, H.O., Paulsen, E.B., Bjerkehagen, B., Szuhai, K., Serra, M., Schaefer, K.L., Myklebost, O., and Meza-Zepeda, L.A. (Aug 2009). LSAMP, a novel candidate tumor suppressor gene in human osteosarcomas, identified by array comparative genomic hybridization. Genes Chromosomes Cancer *48*, 679–693.
37. Wu, G., Yan, Y., Wang, X., Ren, X., Chen, X., Zeng, S., Wei, J., Qian, L., Yang, X., Ou, C., et al. (2019). CFHR1 is a potentially downregulated gene in lung adenocarcinoma. Mol. Med. Rep. *20*, 3642–3648.
38. Fontanari Krause, L.M., Japp, A.S., Krause, A., Mooster, J., Chopra, M., Muschen, M., and Bohlander, S.K. (2014). Identification and characterization of OSTL (RNF217) encoding a RING-IBR-RING protein adjacent to a translocation breakpoint involving ETV6 in childhood ALL. Sci. Rep. *4*, 6565.
39. Sengupta, N., Yau, C., Sakthianandeswaren, A., Mouradov, D., Gibbs, P., Suraweera, N., Cazier, J.B., Polanco-Echeverry, G., Ghosh, A., Thaha, M., et al. (2013). Analysis of colorectal cancers in British Bangladeshi identifies early onset, frequent mucinous histotype and a high prevalence of RBFOX1 deletion. Mol. Cancer *12*, 1.
40. Gong, J., Zhou, Y., Liu, D., Huo, J., and Jun. (2018). F-box proteins involved in cancer-associated drug resistance. Oncol. Lett. *15*, 8891–8900.
41. Imaoka, H., Toiyama, Y., Saigusa, S., Kawamura, M., Kawamoto, A., Okugawa, Y., Hiro, J., Tanaka, K., Inoue, Y., Mohri, Y., et al. (2015). RacGAP1 expression, increasing tumor malignant potential, as a predictive biomarker for lymph node metastasis and poor prognosis in colorectal cancer. Carcinogenesis *36*, 346–354.
42. Yoshihara, K., Wang, Q., Torres-Garcia, W., Zheng, S., Vegesna, R., Kim, H., and Verhaak, R.G. (2015). The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene *34*, 4845–4854.
43. Zhou, B., Ho, S.S., Greer, S.U., Zhu, X., Bell, J.M., Arthur, J.G., Spies, N., Zhang, X., Byeon, S., Pattni, R., et al. (2019). Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. Genome Res. *29*, 472–484.
44. Shi, H., and Wang, Z. (2018). Atypical microdeletion in 22q11 deletion syndrome reveals new candidate causative genes: a case report and literature review. Medicine (Baltimore) *97*, e9936.
45. Matejcic, M., Li, D., Prescott, N.J., Lewis, C.M., Mathew, C.G., and Parker, M.I. (2011). Association of a deletion of GSTT2B with an altered risk of oesophageal squamous cell carcinoma in a South African population: a case-control study. PLoS One *6*, e29366.
46. Reinhart, W.F., Reifenberger, J.G., Gupta, D., Muralidhar, A., Sheats, J., Cao, H., and Dorfman, K.D. (2015). Distribution of distances between DNA barcode labels in nanochannels close to the persistence length. J. Chem. Phys. *142*, 064902.

Patterns, Volume *2*

Supplemental information

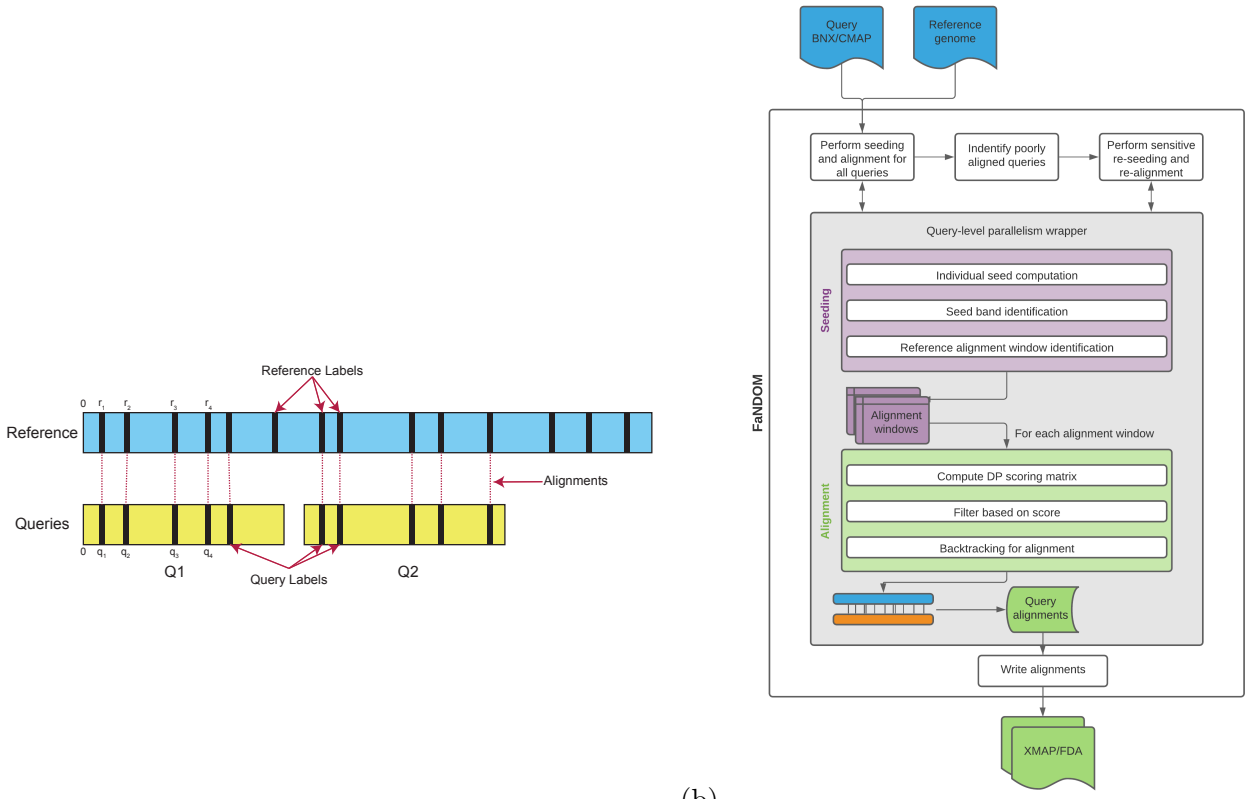
FaNDOM: Fast nested distance-based seeding of optical maps

Siavash Raeisi Dehkordi, Jens Luebeck, and Vineet Bafna

A Supplemental Experimental Procedures

Methods S1 A high-level overview of optical mapping and FaNDOM

FaNDOM utilizes optical mapping data which is converted to BNX or CMAP representations, and outputs optical map alignments in XMAP or its own FDA (**F**an**D**OM **A**lignment) file formats. Figure S1a provides a cartoon representation of multiple optical map 'queries' aligned to an optical map 'reference' segment. The schema of the FaNDOM aligner and its implementation is described in Figure S1b. The seeding and alignment modules are called by each parallel thread to produce and store alignments of query to reference.



(a) (b)

Figure S1: **Overview of OM alignments and FaNDOM software.** (a)Cartoon diagram of optical map queries aligned to *in-silico* reference map. $r1, r2, ...$ and $q_1, q_2, ...$ all represent numeric values in base-pair units of the expected or measured locations of labels in the reference and query, respectively. (b) Overview of FaNDOM software.

Methods S2 Molecule filtering and reference pre-processing

Lower bounds for molecule size filtering were selected using guidance from Bionano documentation available at the Bionano genomics website, page 8. Thus by default FaNDOM ignores molecules < 150 kbp or < 10 labels.

Given that the vast majority of distances between consecutive labels in the reference genome is $\ll 250$ kbp, molecules containing unlabeled stretches > 250 kbp are likely chimeric molecule fragments or incompletely labeled molecules and thus FaNDOM does not attempt to align molecules with unlabeled stretches exceeding 250 kbp.

Optical mapping data also suffers from an error modality in which labels located nearby on a molecule are indistinguishable and measured as a single fluorescent label[15]. FaNDOM pre-processes the reference genome to identify such sites. By default, FaNDOM selects a threshold of 800 bp for which to merge consecutive reference labels (creating an artificial label in the center). The basis for this choice is both theoretical and has some practical support as demonstrated in Luebeck et al., 2020[22]. The approximate length of a single basepair of DNA is approximatel 0.34 nm. The approximate wavelength of green light (used by the label fluorescence laser in the Bionano Saphyr) is approximately 550 nm. This suggests 500 (nm) $\div$ 0.34 (nm x bp^{-1}) $\approx$ 1600 bp of DNA are spanned inside the wavelength of green light. Applying the Abbe diffraction limit of $\lambda/2$ implies a theoretical resolution limit of 800 bp for a given label.

Methods S3 Determining scaling and stretch factors

We define 'scaling' and 'stretch' as two independent error modalities which we examined when benchmarking FaNDOM. *Scaling* refers to the calibration of measured basepairs per pixel in the imaging of optical map molecules by the instrument. If this calibration is not completely accurate, we observed that this error can lead to global lengthening or shortening of all molecules derived from the instrument. To ensure that optical map data has been properly scaled following the image processing performed by the Bionano instrument, we apply a grid-search method to try a range of re-scaling factors. *Stretch* on the other hand refers to the physical lengthening or shortening of individual DNA fragments traveling through the nanochannel array. It is accounted for after 'scaling' has been resolved.

A scaling correction (adjustment of base-pairs per pixel) may need to be applied to data-sets to improve alignment quality. To test how scaling profiles varied across samples, we obtained a a selection molecules from 38 human samples provided by Bionano Genomics. For each sample, we sampled 250 molecules over

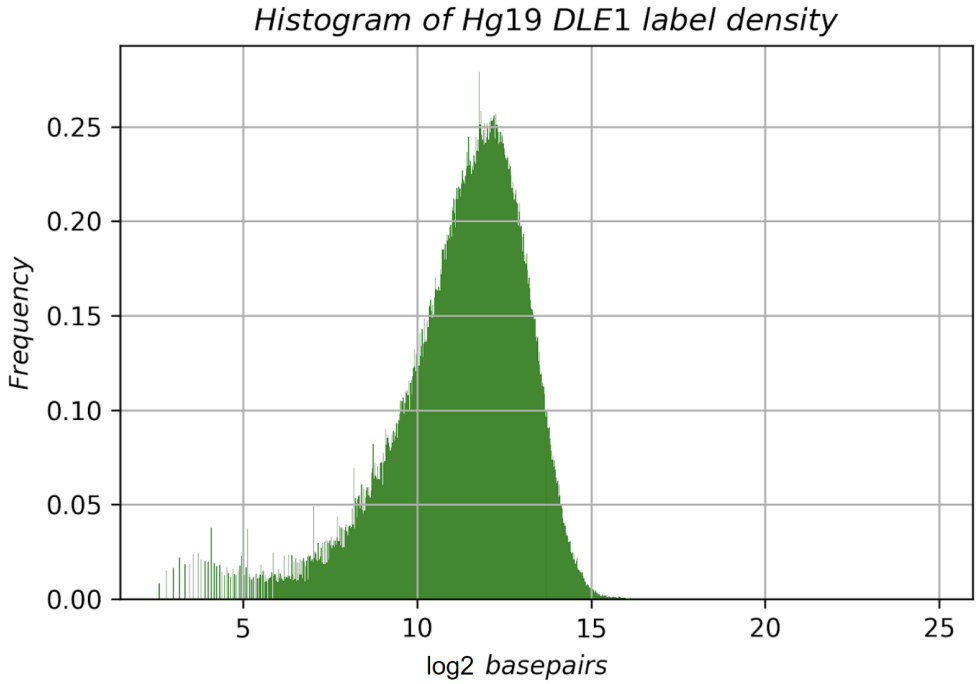


Figure S2: **Histogram of Hg19 DLE1 label density.** DLE1 label density (distance between consecutive label sites) in the hg19 reference genome. Note that the vast majority of the distances are lower than 250kbp $\simeq 2^{18}$bp.

a range of scaling factors from -0.96 to 1.2, and selected a single scaling factor that had the highest sum of alignment scores for all queries. Figure S3a shows the distribution of the best scaling-factors over the 38 samples.

To test for stretch, we selected 100,000 high confidence alignments of molecules. For each alignment of consecutive labels in the query to labels in the reference given by $(q_a, r_b), (q_c, r_d)$, we computed $\frac{|q_a - q_c|}{|r_b - r_d|}$ as the stretch factor. Fig. S3b, S3c shows the distribution of median of stretch factors for 100,000 molecules in two distinct cell-lines with different scaling factors, and suggests a low standard deviation of 0.02.

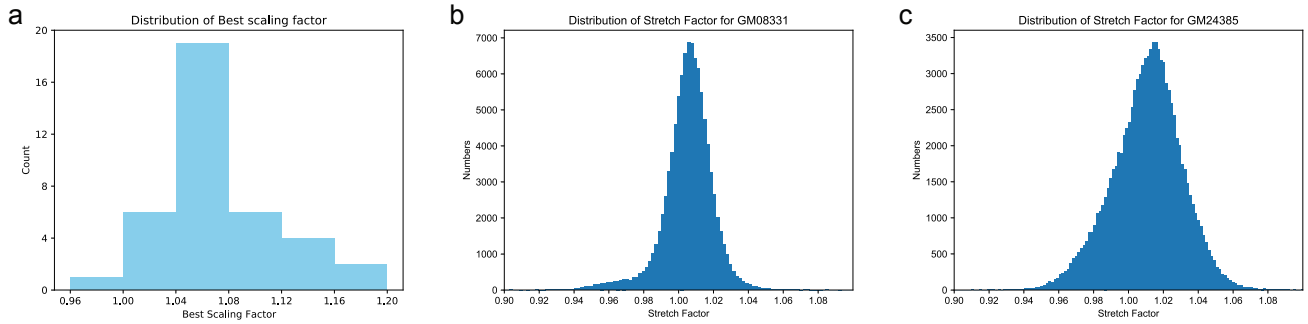


Figure S3: **Scaling factor variation**. (a) Distribution of the single scaling factors for each of 38 human samples achieving highest sum of alignment scores. (b-c) Distribution of estimated "stretch-factors" for samples GM08331 (b) and GM24385 (c) after applying scaling correction.

Methods S4 Setting default values for FaNDOM parameters

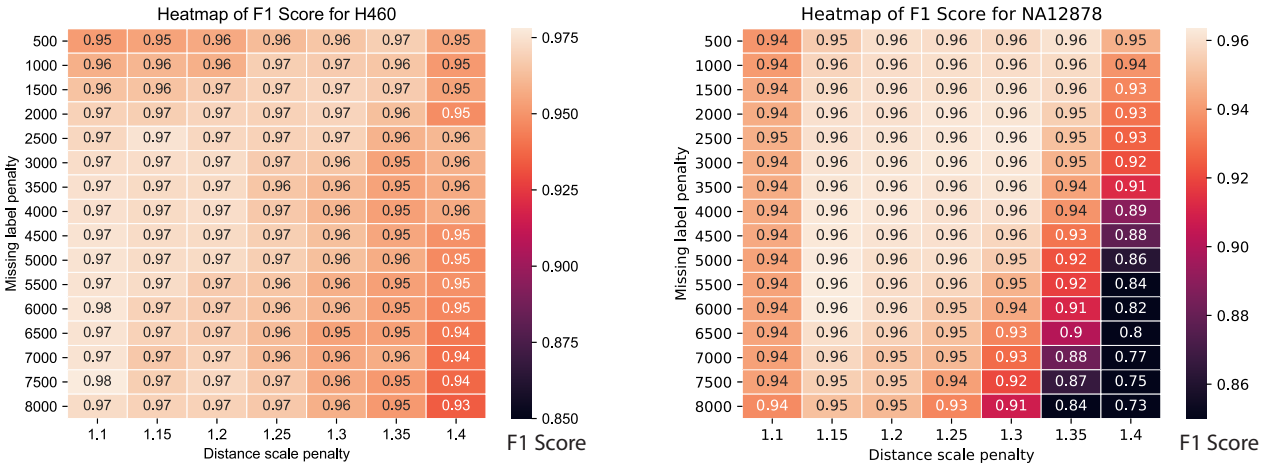


Figure S4: **Alignment score parameters.** Effect of missing label penalty (c) and distance scale penalty (k) parameters on alignment F1 score of (left) H460 OM assembled contigs and (right) NA12878 raw molecules. The performance remains robust for a wide range of parameters.

FanDOM uses a number of parameters that, while modifiable by the user, were optimized for Saphyr technology using a subset of 10,000 alignments computed by the Bionano Ref-Aligner. These alignments were chosen at random but with a small bias towards lower scoring alignments to ensure that we were not optimizing only for the high-quality alignments. The experiments leading to the default settings of the parameters are described below. We first experimented to get crude pre-optimized values for each parameter,

and then optimized each parameter in turn, while keeping the others fixed. We note that FanDOM was robust to small changes in parameters and that the final testing using these parameters on a much larger set and across multiple cell-lines and simulations. Our choice of default parameters was based on achieving maximum speed-up while maintaining recall above 90% for this hard data-set.

Alignment score parameters. We considered the impact of simultaneously varying c (false-label penalty) and k (distance scale penalty) on the F1 score which combines precision and recall. We randomly selected 500 assembled OM contigs of H460 cell-line and 1000 raw molecule of NA12878 cell-line for the tests. Fig S4 indicates that a wide range of k, c showed identical performance. Increasing k, c resulted in the same alignments but with tighter boundaries. We chose the distance scale parameter $k = 1.15$ and false-label parameter $c = 3000$ as default values.

Tolerance. T represents tolerance for matching two genomic distances between query and reference. Increasing the tolerance would increase sensitivity, but would result in increased running time. To address this trade-off, we experimented with T ranging from 200 to 800, plotting running time versus recall. T ranging between 350 and 400 reached our target sensitivity but provided very high speedups. We chose $T = 350$ as default (Fig. S5a). Note that relaxing T to be very large can lead to false alignments and reduced sensitivity.

Width of alignment band. We experimented with $B_w \in [6, 18]$kbp (Fig. S5b). The choice of $B_w = 12$kbp reached our target sensitivity of 90% while maintaining speed of search.

Minimum number of seeds, T_h in a band. T_h represents the minimum number of seed matches within a band for it to be selected. Experimenting with parameters $T_h \in [3, 7]$ provided $T_h = 4$ as the value that achieves 90% recall with high speed (Fig. S5c).

Number of alignments computed. For each of the $10,000$ queries, and each band where a seed threshold was met, we computed a band-score as described in methods, and kept the top 150 band-scores for each query. when we computed the band-score ranks of the true alignments of each of these queries, we observed a range of ranks. For example, in 8417 of 10,000 queries, the true alignment also had the highest band-score and a rank of 1. Nevertheless, the ranks had a long-tailed distribution. We fit the band-scores to an exponential distribution with parameter λ. Fig. S5d plots the rank of the true alignment of a query versus the maximum value of (Band-score/λ) for that query. Note that when the maximum band-score exceeds 2.2λ, the correct alignment is ranked within top 10 in most cases(rectangle with x, y intercepts as $10, 2.2\lambda$ in Fig. S5d). The shaded region in Fig. S5d were chosen to set the cut-offs as below. The points in the un-shaded region corresponded to missed true alignments and represented 3% of the queries.

$$\max_a \text{score}(B_a) \begin{cases} > 2.2\lambda & \Rightarrow \text{Align top 10 bands} \\ > 1.7\lambda & \Rightarrow \text{Align top 50 bands} \\ > 1.5\lambda & \Rightarrow \text{Align top 100 bands} \\ \text{otherwise} & \Rightarrow \text{Align top 150 bands} \end{cases}$$

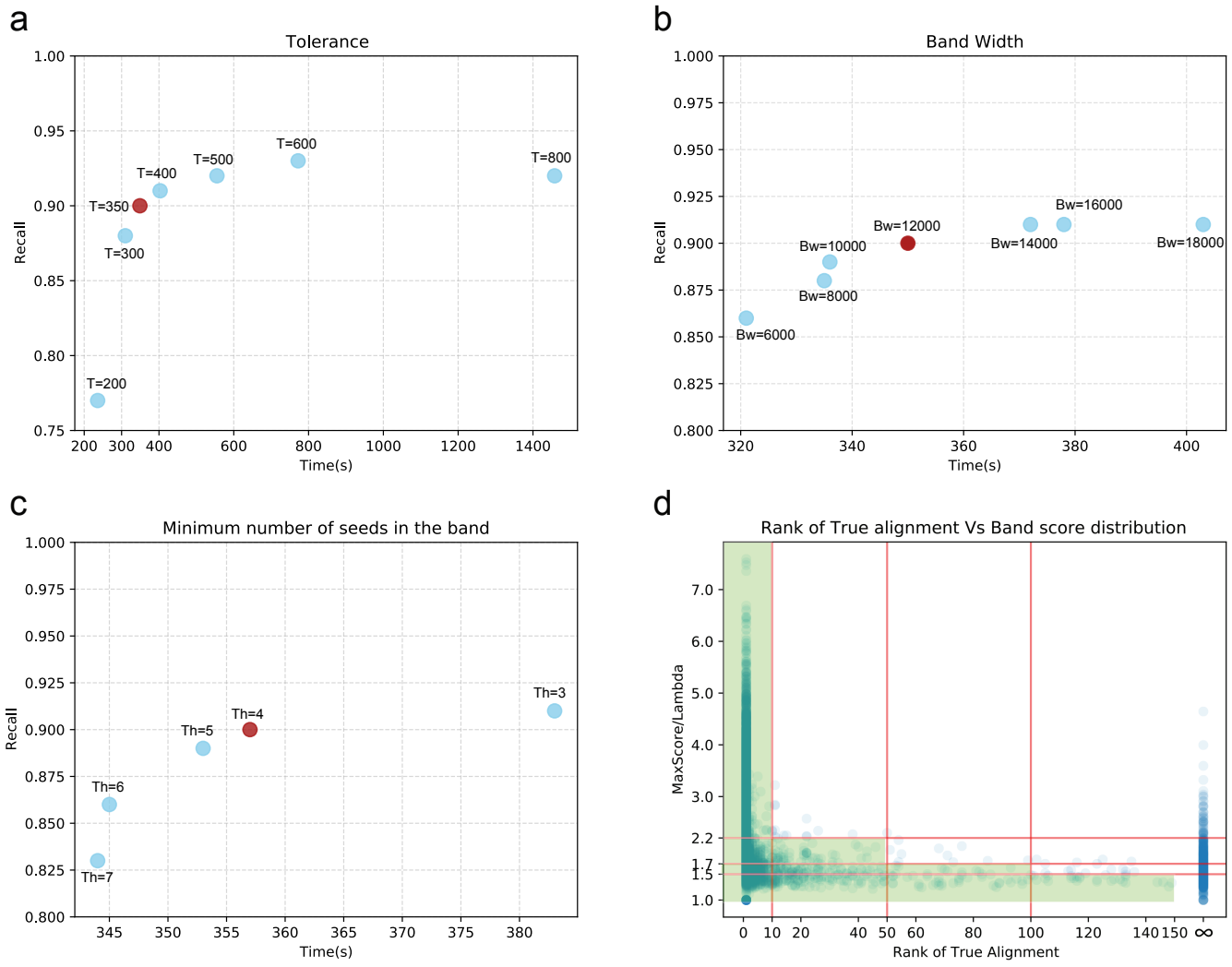


Figure S5: **Parameter tuning.** (a) Tolerance. (b) width of alignment band. (c) Number of seeds. (d) lambda

Methods S5 Experimental framework

All experiments were run on an Intel(R) Core(TM) i9 -9900 CPU @3.10GHz with 32 GB of main memory running Ubuntu 18.04.3 LTS (Bionic Beaver), using 10 threads. The executable file for OMBlast was downloaded directly from from GitHub page and executable file for RefAligner was used from Bionano Solve pipeline version Solve3.5.1_01142020. The following commands were used for each aligner. All aligners were run with default alignment parameters.

FaNDOM:

```
./FaNDOM -t=10 -no_partial -r=ref.cmap -q=query.bnx -sname=output/out
python PythonScript/filter_individual.py -i output/out.xmap
-o output/out_filtered -r ref.cmap
```

OMBlast:

```
java -jar OMTools.jar OMBlastMapper --refmapin ref.cmap --optmapin
query.bnx --optresout OMBlast.xmap --writeunmap false --multiple false
```

```
--thread 10 --minsig 10 --minsize 25000 --minconf 0
```

RefAligner:

```
RefAligner -i query.bnx -ref ref.cmap -o output/out  -maxthreads 10
```

Methods S6 RAM Usage and Running time

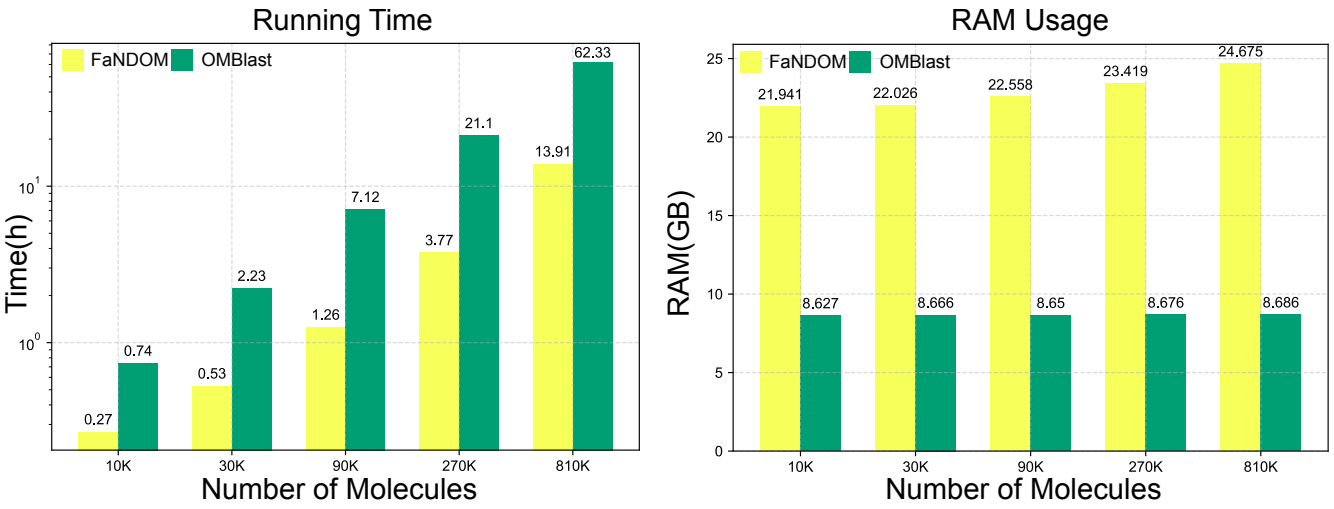


Figure S6: **Scalability of RAM usage and running time**. (left) Running time; (right) RAM usage.

RAM usage of FaNDOM was dependent on the density of labels in the input. Based on the empirical experiments results for Saphyr technology, FaNDOM needed $\sim$ 2.3 GB of RAM per thread on average irrespective of the query molecule length. The memory requirement of OMBlast increased with query size, requiring 200Gb for Saphyr contigs, as per their own documentation.

Methods S7 Simulation

The reference genome used for generating simulated data was hg19_DLE. Genome size was 3095.7(Mbp). The total number of labels was 527922 and average distances between labels was 5,704bp. From the reference genome, 10000 molecules were simulated and extracted by using OMTools simulation[17]. Based on the Bionano documentation for the Saphyr technology, the false positive label rate was 4 per each 100 Kbp and false negative label ratio was 0.1. The average molecule size was set to 250 Kbp. Also based on our calculation of the stretch factor variation (Fig. S3b, S3c), we set the stretch factor standard deviation to 0.02.

High error data:

```
java -jar OMTools.jar OptMapDataGenerator --refmapin hg19_DLE_mas
ked.cmap --flbound 150000 --moleno 10000 --optmapout q.cmap -fsize
250000 --rsln 100 --meas 100 --subound 1.02 --slbound 0.98
--scalesd 0.02 --fpr 0.00004
```

Low error data:

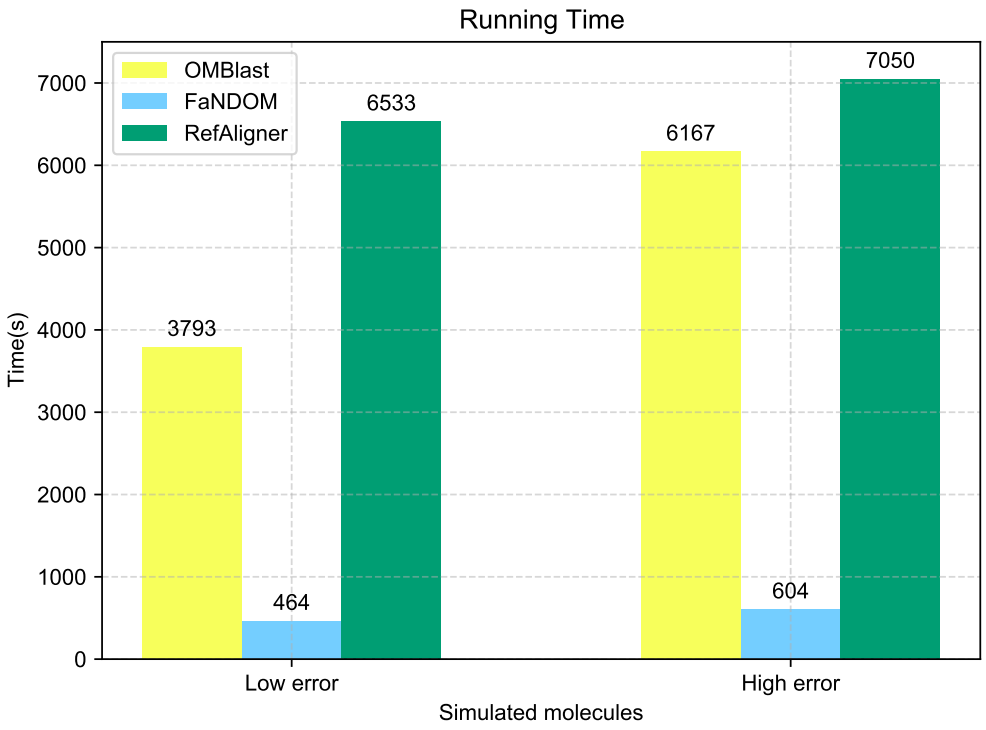


Figure S7: **FaNDOM performance on simulated data.**

```
java -jar OMTools.jar OptMapDataGenerator --refmapin hg19_DLE_mas
ked.cmap --flbound 150000 --moleno 10000 --optmapout q.cmap -fsize
250000 --rsln 100 --meas 100 --subound 1.02 --slbound 0.98
--scalesd 0.01 --fpr 0.00001
```

Methods S8 Comparison against Twin and Kohdista

Much algorithmic work on optical mapping data is related to aligning optical map reads to assemble large genomic optical map scaffolds. Computational tools have been designed for fast identification of overlapping pairs and assembly (e.g. MalignerIX[25], Kohdista[27]). The reference optical map scaffolds can be used for physical mapping of genomic sequences, by *in silico* digestion of the genomic sequence and mapping to the OM reference using tools such as TWIN[26].

While these tools were not explicitly designed to compare optical map fragments to an *in silico* digested genomic reference, we nevertheless benchmarked FaNDOM performance against TWIN[26], and Kohdista[27], which had previously shown fast, memory-efficient performance for the tasks they were designed to solve.

TWIN and Kohdista did not support the Bionano Saphyr technology, which is the dominant platform currently, and required for us to write customfile format converters to convert the modern .bnx, .cmap files into older file formats accepted by TWIN and Kohdista.

We took 10,000 OM molecules from NA12878 and used them as queries to align to the *in silico* digested human reference genome. However, it did not return any mappings. We did test that by using an identical sub-molecule from the *in silico* digestion, we were able to match, suggesting that TWIN showed poor tolerance for missing/false OM labels. TWIN results were previously demonstrated on simulated optical

maps and optical map data with error profiles very different from Bionano Saphyr, so it is possible that a change of parameters could have changed the results. However, in personal communication, the authors did not recommend specific parameter settings appropriate for Bionano Saphyr.

We also attempted to use Kohdista for reference based-mapping. Its RAM usage was not optimized for the large human genome reference. Therefore, we tested a small group of 243 optical map molecules to the genomic reference chr10:0-46,272K (46 Mb). Kohdista used more than 130GB of RAM and did not return alignments after 12 hours. Also, When we tried to align this group of molecules to the entire chromosome 10, the RAM usage surpassed 200 GB. Since optical map data-sets frequently contain > 1 million molecules, we concluded that Kohdista was not an appropriate tool for our problem.