Leveraging linked reads to predict extrachromosomal DNA in advanced prostate cancer

Alex Fortuna[1], Mital Bhakta[1], Natalie Fredriksson[1], Rahul Mannan[2], Fengyun Su[2], Rui Wang[2], Dan Robinson[2], Yi-Mi Wu[2], Xuhong Cao[2], Arul M. Chinnaiyan[2], J. Zachary Sanborn[1], Lisa Munding[1]

1. Dovetail Genomics, Part of Cantata Bio, LLC. Scotts Valley, CA, USA
2. University of Michigan. Ann Arbor, MI, USA

Introduction

Prostate cancer is a complex disease that affects millions of men each year. Recent research has revealed that the hormone-dependent nature of prostate cancer arises from the high expression and genetic amplification of the androgen receptor (AR), often located on extrachromosomal DNA (ecDNA). Identifying and delineating ecDNA from other modes of amplification using whole-genome sequencing (WGS) data remains challenging due to the complexity of copy number and structural variation in cancer. This is compounded in FFPE where this sample type is often at risk of nucleic acid fragmentation, DNA crosslinks, and deamination further complicating variant discovery.

In this study, we investigated if structural variants and linkage patterns derived from Dovetail® DTG-FFPE, an MNase-HiC-based method compatible with FFPE, could predict the presence of ecDNA. We benchmarked our approach using well-characterized cancer cell lines with known ecDNA or HSR, then applied this technique to clinical metastatic castration-resistant prostate cancer (mCRPC) cases.

Methods

To develop our method, we sequenced the COLO320 (ecDNA and HSR variants) cell lines using Dovetail® LinkPrep™ technology. For FFPE mCRPC samples, we generated Dovetail® DTG-FFPE libraries and sequenced to ~30x mean coverage. We utilized BWA for alignment, Hi-C breakfinder for structural variants, DeepSomatic for SNV and InDel detection, Purple for CNVs, and an in-house analytical pipeline for ecDNA detection. AmpliconArchitect was used on orthogonal shotgun WGS to predict ecDNA.

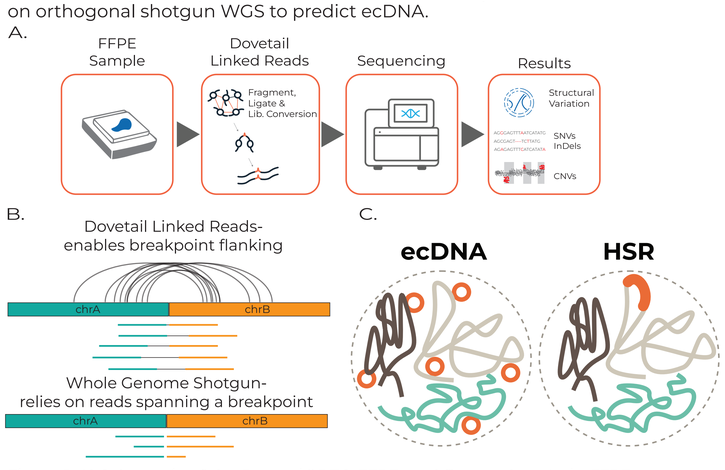

Figure 1: A) Schematic of the Dovetail® DTG-FFPE workflow and analysis. FFPE samples are de-paraffinized, chromatin is conditioned, fragmented with MNase, and religated through proximity ligation. Purified DNA is converted to linked-read libraries compatible with next generation sequencing (NGS). The sequenced libraries are processed through the Dovetail Analysis Portal (https://portal.cantatabio.com). **B) Comparison of SV detection methods between conventional WGS and linked-read sequencing.** Dovetail linked read libraries capture long-range genomic linkage and support high sensitivity detection of SVs. **C) Theoretical model of ecDNA and HSR location within the nucleus.** ecDNA (rings) are free-floating circular structures independent of chromosomes. HSRs are large repeats within a chromosome.

Data Quality

Sample	Read Pairs (Millions)	Duplicate Rate	Long-Range Linkage	Estimated Purity	Estimated Ploidy
COLO320-DM	571	12%	49%	99%	2.6
COLO320-HSR	580	14%	52%	100%	2.6
MI-2	535	13%	57%	64%	2
MI-7	728	14%	42%	39%	2.9

Hypothesis

Samples harboring ecDNAs are identifiable through amplified copy number, increased long range interactions (trans-to-cis ratio), but without a long distance decay pattern typical of a structural variants. Homogeneously Staining Regions (HSRs) have elevated copy number, without increased long range interactions (trans-to-cis ratio), and with the long distance decay pattern.

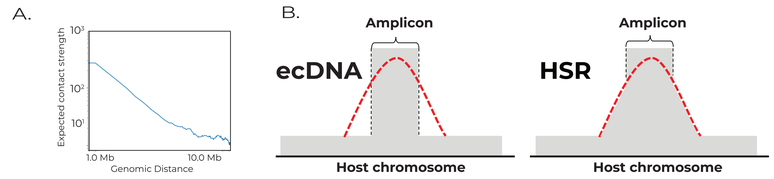

Figure 2: A) Expected counts by genomic distance for linked reads. B) Model illustrating contact frequency decay which allows for distinguishing between an ecDNA derived amplification and HSR.

Results

Methods development in COLO320 cell lines

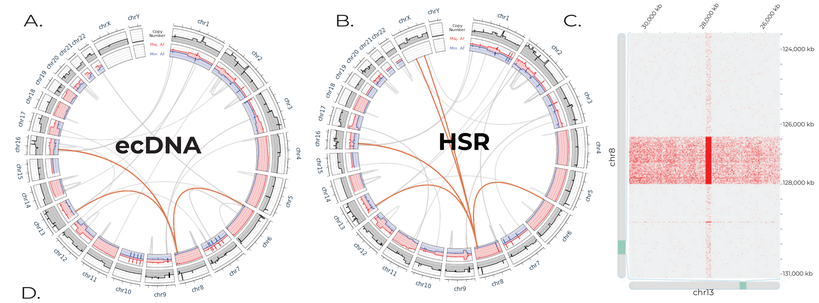

Figure 3: Analytical method development in COLO320 cell lines. A) Circos plot for COLO320-DM and (B) COLO320-HSR. Orange lines highlight structural variants connected to the *MYC* amplicon, including structural variants to two locations on chrX, unique to COLO320-HSR. **C) Contact map highlighting t(8;13). D) COLO320-DM Z-score diagram highlighting significant interactions between amplified segments (> 10 copies). E) COLO320-HSR Z-score diagram highlighting significant interactions between amplified segments (> 10 copies).** Note the significant interaction between amplicon segments and chrX with decay pattern comparable to a structural variant.

mCRPC clinical case study

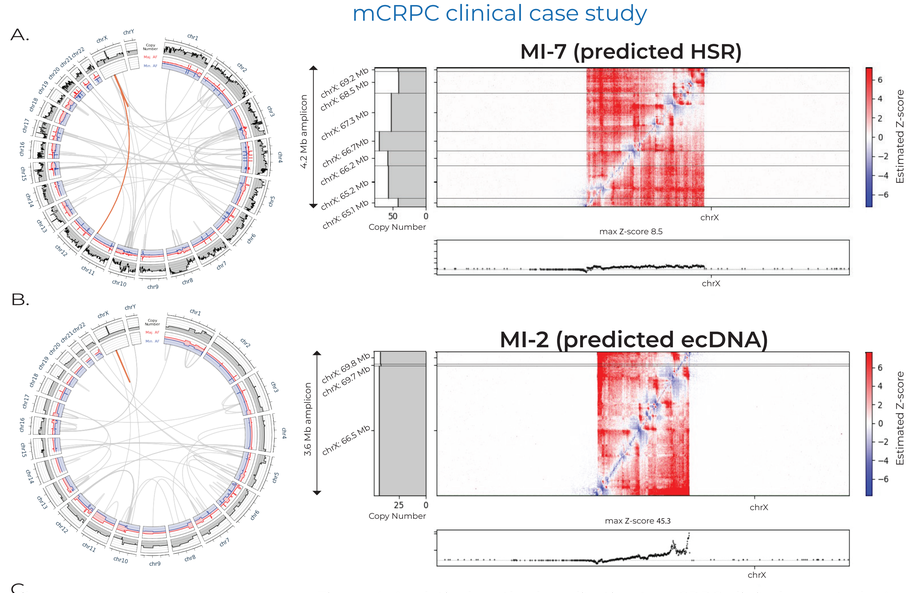

Figure 4: Analytical method application to mCRPC clinical case study. A linkage decay model is used to calculate Z-scores of observed vs. expected linkage from the 5' edge of the AR amplicon to other positions within the amplicon and the regions of chromosome X directly adjacent to the amplicon's sequence in hg38 for samples **(A) MI-7** and **(B) MI-2**. The sharp increase in Z-score observed at the 3' edge of the AR amplicon in MI-2 is explained by the circular structure characteristic of an ecDNA.
C) AmpliconArchitect reconstruction of the AR-containing amplicon from fresh frozen (FF) whole-genome sequencing of sample MI-2. The analysis identifies a circular extrachromosomal DNA (ecDNA) element spanning ~3.6 Mb on chromosome X which harbors AR. AmpliconArchitect predicts approximately 33 copies of this amplicon, consistent with copy number profiles generated from the DTG-FFPE analysis pipeline.

Sample	30x max Z-score	30x log2ratio	30x Gini Coefficient	10x max Z-score	10x log2ratio	10x Gini Coefficient	5x max Z-score	5x log2ratio	5x Gini Coefficient
MI-2	45.3	-1.12	0.34	25.9	-1.12	0.33	21.6	-1.19	0.34
MI-7	8.5	-2.67	0.34	4.6	-2.65	0.34	3.4	-2.65	0.33

Figure 5: Statistical summary of ecDNA and HSR prediction from in silico downsampling. The max Z-scores is the peak score of observed vs. expected linkage from the 5' edge of the AR amplicon**.** The log2ratio represents the ratio of inter contacts and intra contacts between the amplicon and the rest of the genome. In the predicted ecDNA (MI-2), the max Z-scores are over 5x higher than the max Z-scores of the predicted HSR (MI-7), while also showing an increase in the log2ratio (i.e. relative increase in the number of inter contacts). This pattern is robust to 5x coverage. The Gini coefficient is a measure of contact inequality and there is no noticeable difference between samples.

Conclusions

- Unlike conventional short-read sequencing, Dovetail® linked reads use breakpoint flanking sequences for structural variant detection, enabling higher sensitivity and accuracy by filtering out many false positives.
- We developed an analytical method compatible with Dovetail® linked reads to evaluate ecDNA probability based on genomic linkage. ecDNA/HSR prediction is now available through Variant Analysis on the Dovetail Analysis Portal (https://portal.cantatabio.com)
- Dovetail® linked reads (LinkPrep™) identify known structural segments and their linkage in COLO320 cell lines. Together demonstrating performance in reference material. By using Z-scores of observed vs. expected linkage we predicted the location of the MYC containing amplicon on chrX, absent in the ecDNA cell line.
- In clinically-derived mCRPC samples, Dovetail® linked reads (DTG-FFPE) detect clinically relevant amplifications and predicts whether the amplicons are found in ecDNA or HSR structures within the tumor genome. MI-2 has a strong ecDNA signal (high max Z-score) which is robust to 5x.