

Technical Note

Improved Whole Genome Somatic Variant Discovery

Key Takeaways

- Achieve ultra-high sensitivity structural variant detection with base pair resolution.
- Detect SNVs, InDels, and CNVs with WGS-like capability.
- Detect the full spectrum of somatic variation using a single NGS assay.

Introduction

Genetic variation plays a crucial role in many diseases, especially cancer where mutations drive disease development and progression. Somatic variants – such as single nucleotide variants (SNVs), small insertions or deletions (InDels), copy number variations (CNVs), and large structural variants (SVs) - can disrupt cellular functions by altering gene expression,

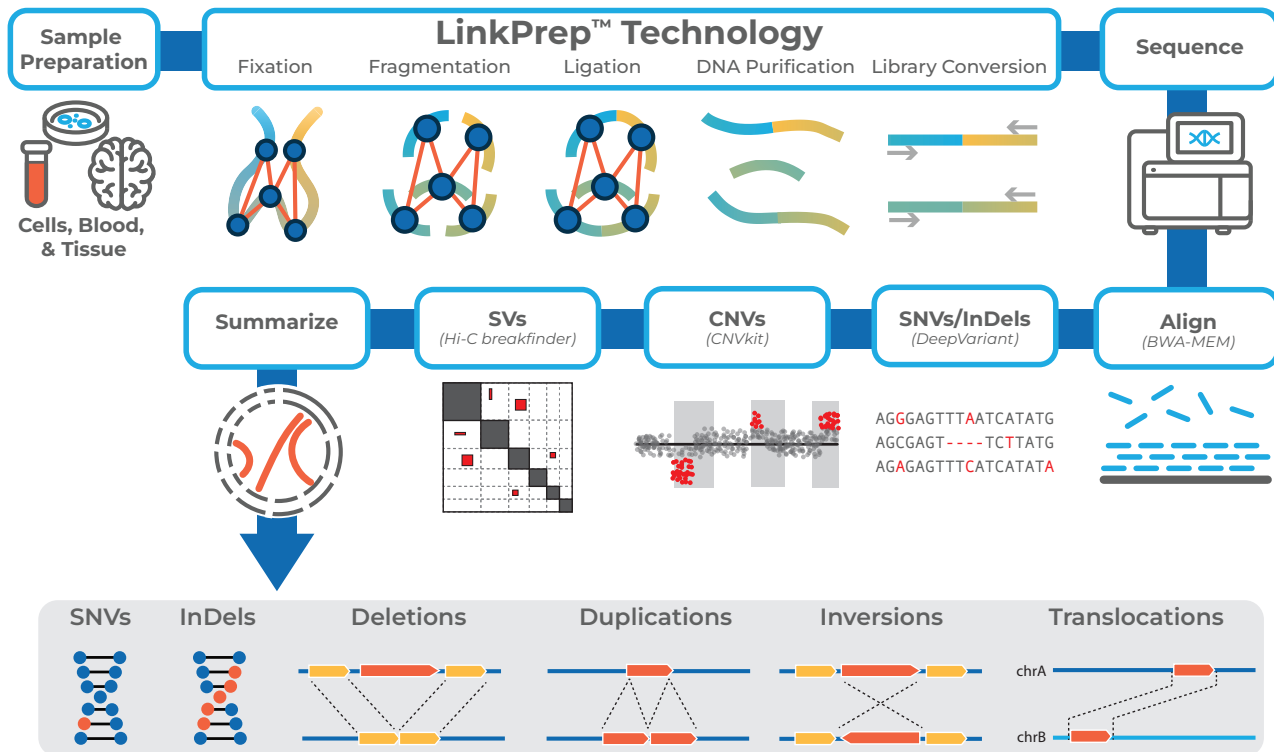


Figure 1. Capture the full spectrum of somatic variation through a single simplified workflow. Samples (cells, tissue, or blood) are processed through the LinkPrep technology and sequenced on a short-read platform. The resulting data are analyzed with standard NGS software for SNVs/InDels (DeepVariant), CNVs (cnvkit), and SVs (Hi-C breakfinder), resulting in a complete catalog of genetic variation contained within a sample.

thereby activating or inactivating oncogenes or tumor suppressor genes, respectively.

Detecting the full spectrum of somatic variation requires multiple assay types due to variability in size and complexity. SV detection is particularly challenging with traditional whole genome sequencing (WGS) approaches as it is hampered by read length and a high sequencing depth needed to capture breakpoint informative reads. It is further burdened by many false positives.

Here we introduce Dovetail® LinkPrep™ technology. The method generates linked-read libraries through DNA proximity ligation. LinkPrep™ data simultaneously provides highly sensitive detection of SVs while retaining WGS capabilities for the detection of SNVs, InDels, and CNVs in a single NGS dataset (Figure 1). The end-to-end assay proceeds with *in situ* sample fixation, fragmentation, and ligation, followed by a coupled library conversion process. The single-

day workflow requires no specialized laboratory equipment, and is compatible with any short read, paired-end sequencer and standard NGS analysis software for variant calling and classification.

High sensitivity SV detection with short reads

The linked-read attribute of LinkPrep data boosts sensitivity for SV detection through breakpoint ‘flanking’, the ability to use the surrounding genomic space to provide evidence of a breakpoint. Uniform per base coverage enables further breakpoint refinement (Figure 2). The combination of these two capabilities makes LinkPrep data ideal for large SV detection (down to 20% tumor fraction/ 10% variant allele frequency), resolving complex structural events, and enhancing detection of homologous recombination deficiency (HRD).

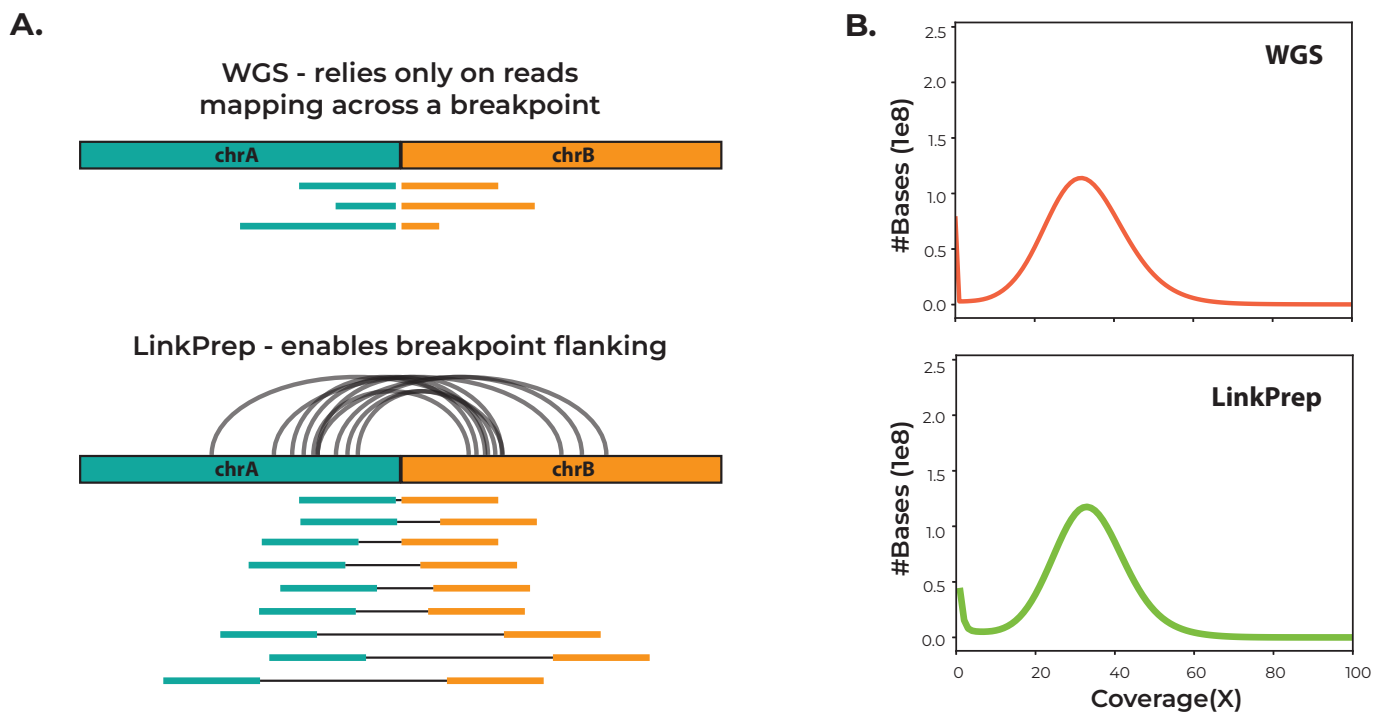
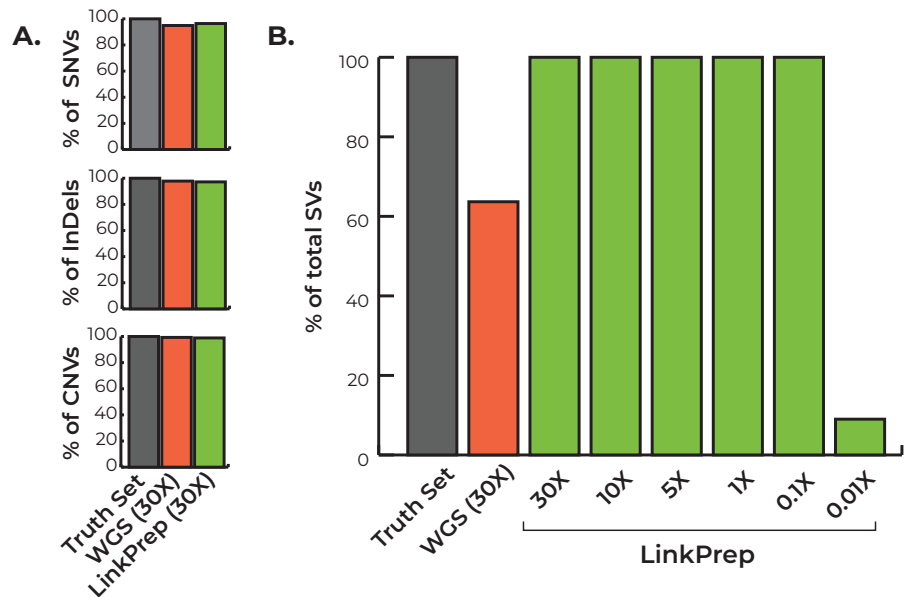


Figure 2. LinkPrep technology bridges the gap between large scale SV mapping and nucleotide-level variant detection. A) Large variant detection using WGS methods (long- or short-read) relies on identifying rare reads that map across the breakpoint providing evidence for a structural rearrangement in the genome. The linked-read attribute of LinkPrep data facilitates the use of breakpoint flanking reads to improve sensitivity for SV detection. B) Since LinkPrep data has a similar per-base coverage distribution as WGS, it is effective for small variant detection.

Figure 3. LinkPrep data excels at detecting the full spectrum of genetic variants. K562 cells were profiled using both the LinkPrep Assay and WGS (sequenced to 30X coverage) and the results compared to a K562 truth set of known variants. The truth set was generated using a variety of technologies including WGS at >80X coverage, RNA-seq, long read sequencing. A) The LinkPrep data demonstrated WGS-like capabilities in detecting SNVs, InDels, and CNVs. B) The LinkPrep Assay demonstrated a much higher sensitivity for SVs compared to WGS. To further demonstrate the extreme sensitivity of SVs, LinkPrep data was down sampled to 0.01X genomic coverage.



Do more with your short reads

Relying on the fidelity of short-read sequencers, LinkPrep technology captures SNVs, InDels, CNVs and large SVs in a single assay. To benchmark performance, the ability of the LinkPrep assay and WGS to detect variants in the well-characterized leukemia cell line, K562 (a truth set of known variants has been defined by a combination of technologies including, WGS at >80X coverage, RNA-seq, and long-read sequencing) was compared.

At 30X genomic coverage, both WGS and LinkPrep data show near perfect concordance with the truth set for SNVs, InDels, and CNVs (Figure 3A). In contrast for SVs, WGS data detected only ~60% of the truth set, whereas LinkPrep data detected 100% of the SVs compared to the truth set (Figure 3B). Importantly, LinkPrep data retained 100% SV recall down to 0.1X coverage (1 million read pairs) in a down-sampling experiment.

Performance in clinical samples

The LinkPrep assay was applied to two well-characterized cell lines (HCC1187 and K562), as well as a series of clinical samples (Figure 4). The clinical samples included two separate ovarian carcinomas (serous adenocarcinoma and metastatic carcinoma) and a nasal structure

osteosarcoma. All samples were sequenced to 30X.

Again, the LinkPrep data demonstrated a high concordance with WGS data for the detection of SNVs, InDels, and CNVs across all samples. For SVs, WGS data analysis reported 100s-1000s of SVs per sample suggesting a high false positive rate. The majority had incredibly low read-support requiring extensive secondary false-positive filtering.

In contrast, SVs detected by LinkPrep data were all highly supported (>500 reads). The false positive difference between WGS and LinkPrep data was reflected in the F1 scores computed for the cell line samples. Furthermore, the top-ranked clinically relevant SNVs and InDels were captured by both assays with a 100% overlap.

LinkPrep data enables variant visualization with standard methods such as Circos plots to display large SVs and CNVs. Improved detection of somatic variants, including large SVs, using LinkPrep technology, enhanced identification of oncogenic drivers that may be missed by WGS or buried in a long list of false positives.

SV breakpoint refinement

Unlike other technologies geared towards sensitive SV detection (e.g. optical genome mapping or FISH), LinkPrep data's uniform

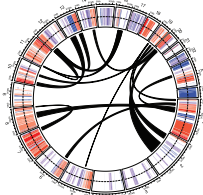
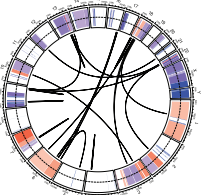
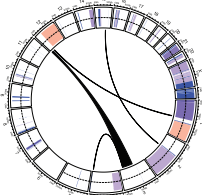
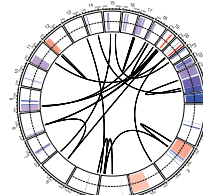
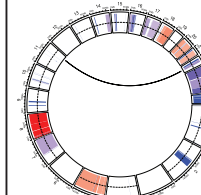
		Cell Lines				Clinical Samples					
Sample Info	Sample	Ductal Carcinoma (HCC1187)		Myelogenous Leukemia (K562)		Serous Adenocarcinoma (Ovarian)		Metastatic Carcinoma (Ovarian)		Osteosarcoma (Nasal)	
	Tumor Fraction	100%		100%		72%		32%		98%	
SVs	Assay	LinkPrep	WGS	LinkPrep	WGS	LinkPrep	WGS	LinkPrep	WGS	LinkPrep	WGS
	#SVs Called	15	2,311	19	1,800	4	26	17	3,361	1	839
	F1 Score	0.938	0.012	0.348	0.006	-	-	-	-	-	-
	Mean Read Support	50,000	43.6	3,328.3	39.3	3,329.9	25.3	530.9	39.6	836	40.4
CNVs	Correlation (> 3 copy number)	95.73%		97.44%		NA		89.35%		100%	
SNV/InDels	Total SNVs	399	453	665	741	381	382	300	483	329	511
	Total InDels	43	36	40	42	20	21	20	26	18	35
	% Overlap	98.09%		98.00%		92.90%		97.98%		98.28%	
	Candidate Pathogenic	2	2	2	2	2	2	2	2	4	4
Summary	LinkPrep Circos (SVs & CNVs)										
	Notable Oncogenes Across All Variant Classes	SHANK3 CCNE1 CCND2 AMPD2 TGDS		BCR-ABL1 ASXL1 EZH2 BRAF ASL ASXL1		MECOM KRAS CHEK2 KIRREL1		HNF1A TP53 PEX1 AKT2 TPP1		ESWRI fusion MYC PREPL NPHP1 PAX8 IYD	

Figure 4. LinkPrep technology outperforms WGS in clinical samples. A side-by-side comparison of variants detected in cell lines and clinical samples. In all samples, LinkPrep data more accurately captures SVs with minimal false positives (calculated by the F1 score for cell lines). Note that HCC1187 has a high confidence truth set, while the publicly available truth set for K562 is less curated and contributes to the lower F1. CNV analysis regressions were performed on all genes amplified greater than 3 copy numbers between WGS and LinkPrep data. Across all samples, LinkPrep™ reads capture SNVs and INDELs with high accuracy, as identified by DeepVariant and detects the same clinically relevant small variants (classified by ClinVar as “pathogenic” or “likely pathogenic” with gnomAD population allele frequencies less than 1%). For each sample, LinkPrep™ results are easily summarized in an industry-standard Circos plot with arcs denoting SV events, and a CNV track with gains and losses indicated in red and blue, respectively. Additionally, notable oncogenes are listed that are linked to the various classes of genetic variation.

coverage enables mapping of the SV breakpoint down to base pair resolution.

Initially, SVs are called at moderate resolution, such as within a 10kb bin size, ensuring that each breakpoint is contained within a single large bin. To resolve the exact breakpoint position, the sequence coverage within the identified bin can be reviewed for fluctuations;

the position at which sequence coverage changes indicates the breakpoint (Figure 5). Knowledge of the exact breakpoint position simplifies validation through PCR or Sanger sequencing.

Sample and sequencing requirements

LinkPrep technology requires intact cells

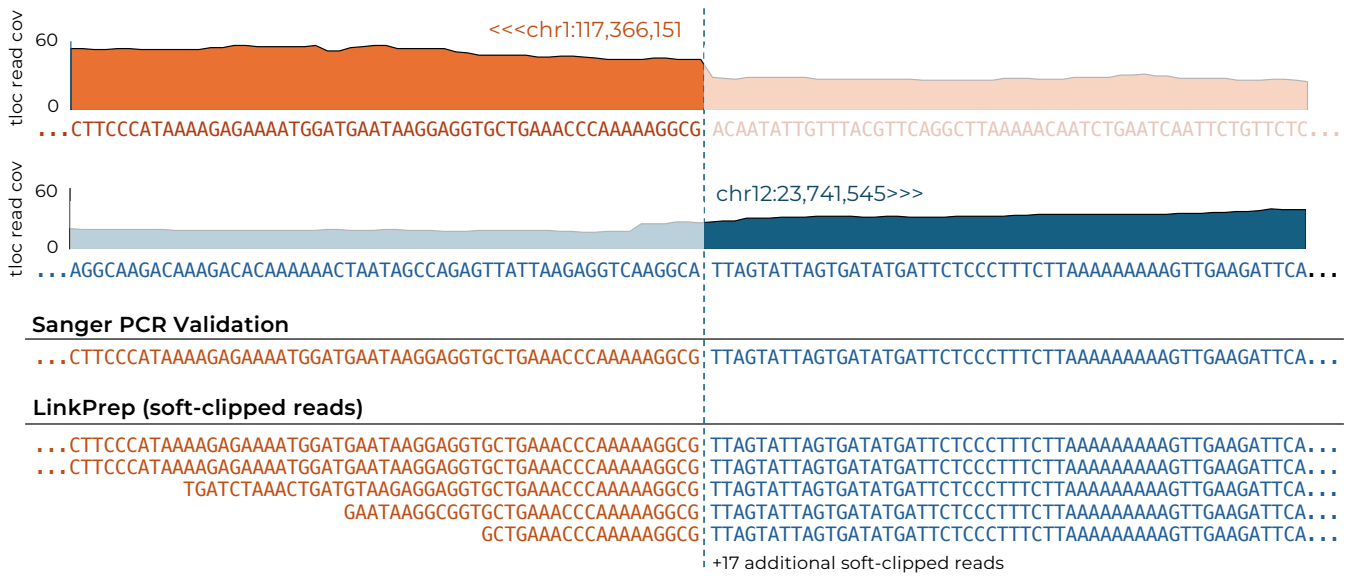
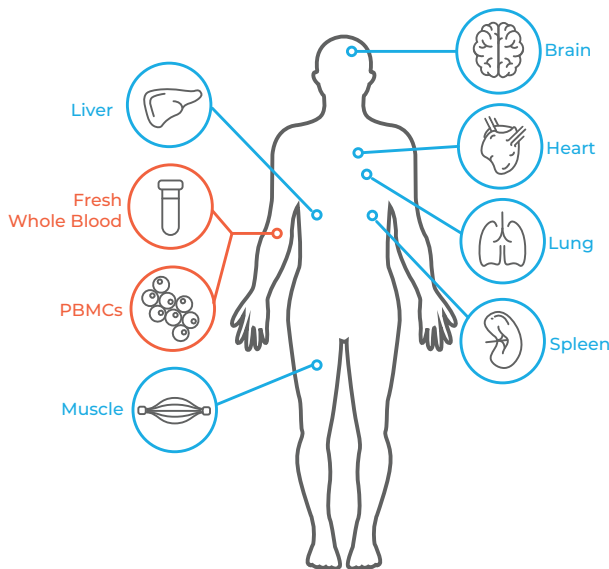


Figure 5. LinkPrep data coverage enables base pair breakpoint refinement. In the ovarian serous adenocarcinoma sample, the breakpoint of the t(1;12) translocation is initially reported within a 100kb window. The breakpoint can be further refined through changes in coverage due to uniformity of the LinkPrep data. A zoomed-in view shows a coverage drop at (or near) the precise breakpoint location for both chr1- and chr12-originating reads (orange and blue, respectively). PCR followed by Sanger sequencing confirms this breakpoint and shows clear alignment with LinkPrep reads, validating the detected SV and its refined breakpoint.

as sample input and is compatible with cells (1 million), flash frozen tissue (3-20 mg dependent on tissue source), whole blood (3 mL), and PBMCs (1 million). Sample purity and the variant types detected influence the

sequencing depth required.

For clonal variants in tumors with >50% purity, 10X coverage (~100 million read pairs) is sufficient for SV detection. Increasing to ≥30X



Tumor Fraction and Sequence Recommendations

Tumor Fraction	Variants Classes	Coverage*	#Read Pairs
>50%	SVs	10X	100 Million
>50%	SVs, CNVs, InDels, & SNVs	30X	300 Million
50-20%	SVs	30X	300 Million
50-20%	SVs, CNVs, InDels, & SNVs	80X	800 Million

*Coverage is based on 2x150bp read pairs
 Detection of low VAF SVs is dependent on sequencing depth

Figure 6. Validated samples, tumor fraction, and sequencing recommendations. A schematic representing a subset of the sample types that have been validated for LinkPrep technology, for a complete up-to-date list please see the LinkPrep User Guide. Not all researchers may be interested in capturing both small and large variants, as such the LinkPrep sequencing requirements provides flexibility enabling users to sequence depending on their needs. Sequencing requirement will be dependent on tumor fraction and the VAF of the respective variant class.

(~300 million read pairs) extends detection to SNVs, InDels, and CNVs. Low frequency variants and/or lower purity samples (20-50% purity) require additional sequencing depth across all desired variant classes (i.e. 30X coverage) for SV and 80X (~800 million read pairs) to detect SNV, InDels, and CNVs; Figure 6).

Conclusion

In summary, the LinkPrep technology provides a robust, efficient solution for comprehensive somatic variant detection, offering high sensitivity for large structural variants while maintaining base pair resolution for small variant detection. Offering a rapid, streamlined workflow compatible with standard NGS platforms, LinkPrep technology enables simultaneous detection of SNVs, InDels, CNVs, and SVs in a single assay. By enhancing breakpoint mapping accuracy and reducing false positives, LinkPrep technology addresses the limitations of traditional WGS, supporting improved insights into complex genetic diseases.



For more information, visit
<https://cantatabio.com/dovetail-genomics/products/dovetail-linkprep-kit/>