High-resolution Sequencing and Graph-based Computational Platforms Enable Routine Cancer Genome Reconstruction

Integration of HiFi and Hi-C sequencing with graph-based approaches reconstruct chromosome-scale structural variant (SV) landscapes for cancer genomes with high precision and accuracy.

The complexities of cancer genetics present a unique challenge for researchers unraveling mechanisms to improve personalized disease diagnosis and management.[1] Characterizing chromosome-scale and haplotype-resolved (i.e., groups of alleles along a single chromosome) mutations in the cancer genome is critical for understanding clonal evolution and cancer progression. However, current methods lack the capability to obtain accurate and precise SV landscapes required for comprehensive cancer genome reconstruction.

While several reports describe the detection of SV events using short-read WGS data,[2] more recent studies have leveraged cutting-edge technologies for precise genome reconstruction, such as long accurate HiFi and long-range Hi-C technologies.[3] Nevertheless, either of these techniques alone is insufficient for cancer genome reconstruction. Novel integrative protocols are required.

Challenge: Developing Efficient Protocols for Chromosome-scale Genome Reconstruction

Although methods exist for chromosome-scale and haplotype-resolved genome reconstruction, they have several disadvantages that limit their application for cancer genomes and lack the necessary throughput for routine analyses. For example, high-resolution HiFi sequencing has effectively produced long reads at the chromosome level but has not been applied to cancer genomes. Similarly, several widely used analytical tools, such as HiCanu, hifiasm, and Flye, are not designed for detecting chromosome-level events in cancer genomes. Other genome reconstruction techniques have been employed but require trio information that is not routinely available.

To address these pitfalls, Shilpa Garg at the Technical University of Denmark and the University of Copenhagen integrated HiFi and Hi-C sequencing with a graph-based algorithm called "pstools" to achieve chromosome-level and haplotype-resolved cancer genome reconstruction.

Approach: Integrating HiFi and Hi-C Sequencing Data Using a Graph-Based Algorithm

To reconstruct cancer genomes at the chromosome level, Garg sequenced the melanoma COLO829 cancer cell line with HiFi and Arima Hi-C sequencing. The PacBio reads were processed using pstools, an approach that combines multiple data types in a joint sequence space that preserves genome complexity, provides haplotype information, and produces complete phased genomes.[4]

Technology

The Arima Genome-wide HiC+ kit provides 3D genome mapping with high uniformity to enable long-range analyses of variant discovery, base polishing, scaffolding, and phasing.

Pstools is a graph-based algorithm that first produces a sequence graph at the whole-genome level that preserves haplotype information in overlapping HiFi read sequences. Hi-C reads are then introduced through the HiFi sequence graph to produce haplotype paths and connect phased contigs in the correct order and orientation (a process known as haplotype-aware scaffolding; see Figure 1). In contrast to competing methods with run times of >2 days, this workflow is highly streamlined and can be completed in less than 12 hours with high accuracy.

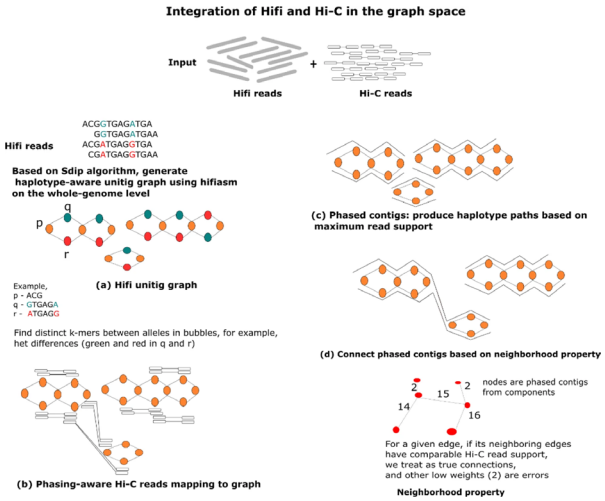


Figure 1. Workflow for processing HiFi and Hi-C reads with pstools, a graph-based algorithm for chromosome-level genome analyses.[4]

This protocol was benchmarked against publicly available SV call sets that utilized multiple technologies, including the short-read-based NYGC (New York Genome Center), single-cell–based, and UMCU (University Medical Center Utrecht) call sets. Notably, the results showed that the calls made with pstools agreed with those made with single-cell and UMCU callsets (Figure 2), with the added advantage of higher precision in base and haplotype resolution SV characterization.

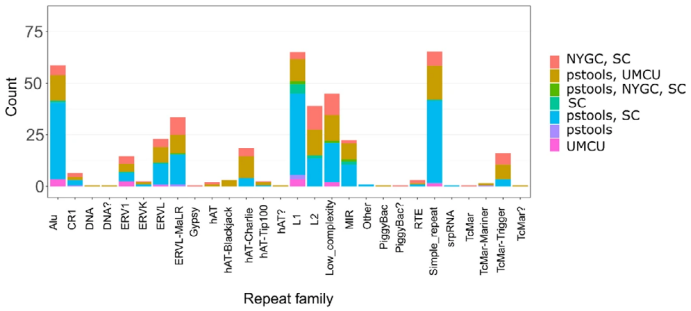


Figure 2. Benchmarking against single-cell (SC), NYGC, and UMCU call sets demonstrates agreement in SV discovery, with each bar displaying the number of variants agreed between call sets for each repeat element.[4]

The integrative protocol presented by Garg provides two key advantages:

- It correctly disentangles chromosomes by connecting components using long-range Hi-C
- It correctly, accurately, and precisely generates phased scaffolds

As shown in Figure 3, pstools utilizes Hi-C information (chr13 and 14; red and green paths) to accurately connect the starting regions of chromosome arms (chr1 and 6). This provides substantial utility for routine clinical applications, particularly when examining inter- and intra-chromosomal structural sequence events.

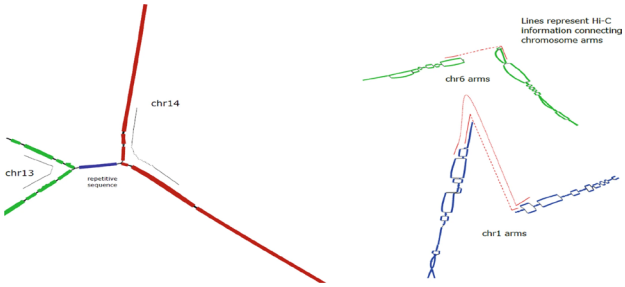
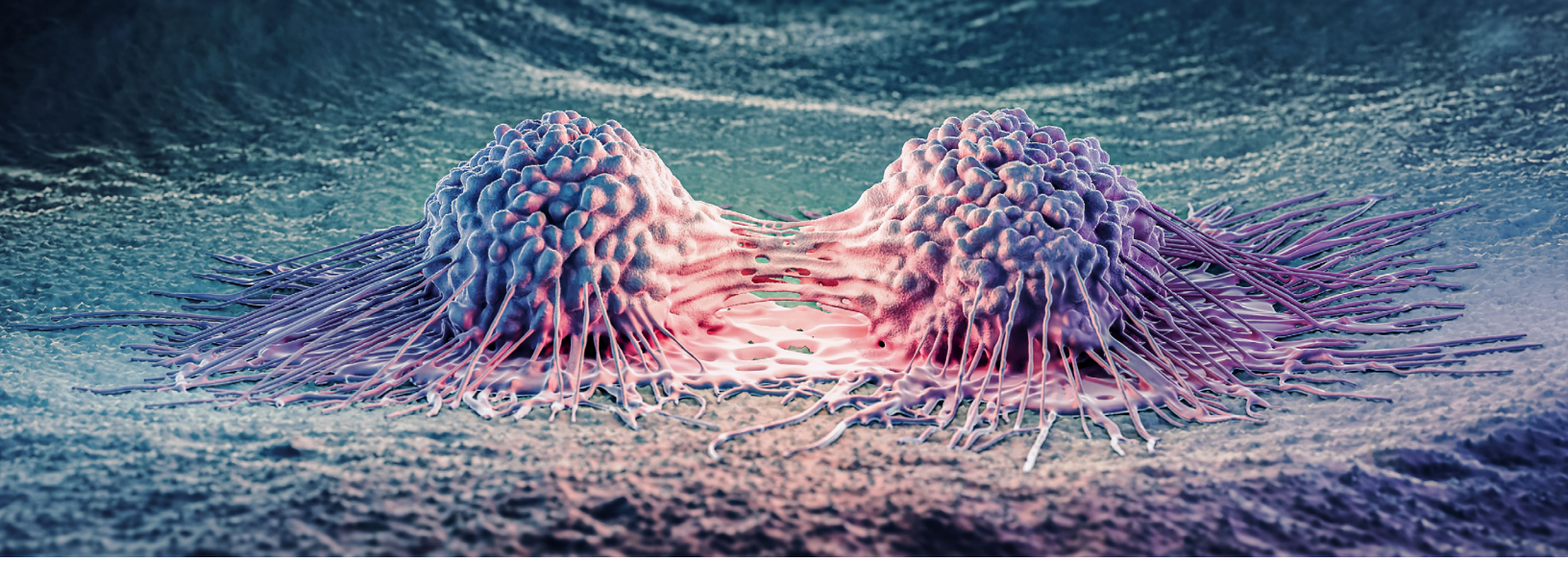


Figure 3. Integration of high-resolution sequencing and pstools facilitates accurate reconstruction of chromosome-scale haplotype-resolved genome assemblies. Hi-C information connects chr13 and chr14 (left) and correctly connects arms of chromosomes (right) with phasing information for Hi-C.[4]

To further demonstrate the utility of these protocols, Garg also benchmarked the pstools algorithm against healthy human samples (HG002, HG00733, PGP1). Trio data were obtained from trio-hifiasm contigs and/or salsa2 for scaffolding. Interestingly, competing methods such as hifiasm (Hi-C) produced assemblies with sizes of NG50 <52Mb, demonstrating that these techniques are not designed for chromosome-scale genomics.

While introducing trio data (trio-hifiasm) improves assembly size, pstools outperforms these methods with high-quality assemblies, producing NG50 assembly sizes of >130Mb and scaffold sizes of >6Gb. Although the addition of a salsa2 Hi-C scaffolding step for trio-hifiasm improved assemblies for this protocol, the pstools approach yielded comparable phasing accuracy with faster runtimes without trio information.

Finally, pstools can uncover precise chromosome-level SV landscapes. The workflow identified 19,956 insertions, 14,846 deletions, 421 duplications, 52 inversions, and 498 translocations in the COLO829 cell line. Analysis of existing multi-technology, high-confidence SV call sets revealed an f1 score of



93.9% achieved by pstools compared to <82% achieved by Hifiasm+salsa2 and Hifiasm+3D-DNA. Additionally, pstools facilitated the discovery of multi-event SVs, such as breakage-fusion-bridge events. Together these results suggest that the precision, accuracy, and efficiency of pstools are superior to existing methods.

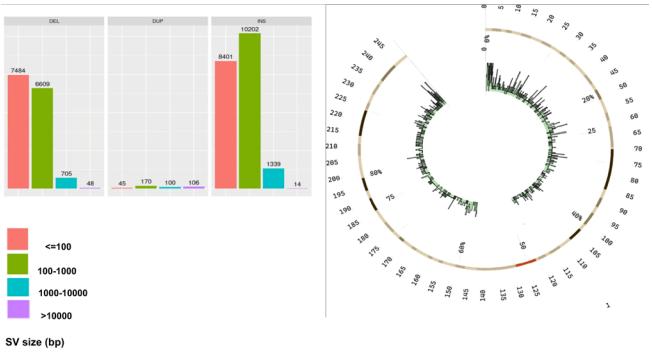


Figure 4. Pstools uncovers chromosome-level SV landscapes with high precision. Charts display SV calls (left) and SV distribution (right).[4]

Impact: Efficient Tools Facilitate Personalized Insights

In the present study, Garg describes an integrative protocol that enables chromosome-scale and haplotype-resolved cancer genome reconstruction. Because most current HiFi and Hi-C technologies alone are insufficient for chromosome-scale cancer genome reconstruction, Garg developed a pstools approach that utilizes the advantages of both HiFi and Hi-C to reconstruct the SV landscape of cancer genomes precisely and accurately. In addition, the fast runtime of this approach increases throughput and facilitates routine analysis for numerous applications.

With the complexity and heterogeneity present in cancer genomes, innovative technologies, and tools are required to gain a comprehensive view of SV landscapes. This is critical for efficiently analyzing hundreds to thousands of diverse samples. Furthermore, with higher-throughput production of fully phased sequences at the chromosome scale, the approach described in this study may facilitate improvements in personalized diagnosis, disease management, and new biological discoveries.

References

1. Wang, W.-J., et al. (2020). Chromosome structural variation in tumorigenesis: Mechanisms of formation and carcinogenesis. *Epigenetics & Chromatin, 13*(1), 49.
2. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, et al. (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*(7793), 82–93.
3. Garg, S. (2021). Computational methods for chromosome-scale haplotype reconstruction. *Genome Biology, 22*(1), 101.
4. Garg, S. (2023). Towards routine chromosome-scale haplotype-resolved reconstruction in cancer genomics. *Nature Communications, 14*(1), 1358.

Learn More Webinar: Where Will Genomes Take Us Next: How Chromosome-Scale Assemblies Are Unlocking New Biology