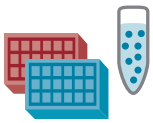
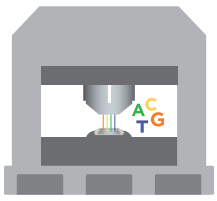
Revolocity™ Whole Genome Sequencing Technology Overview

DNA EXTRACTION



LIBRARY PREPARATION



SEQUENCING

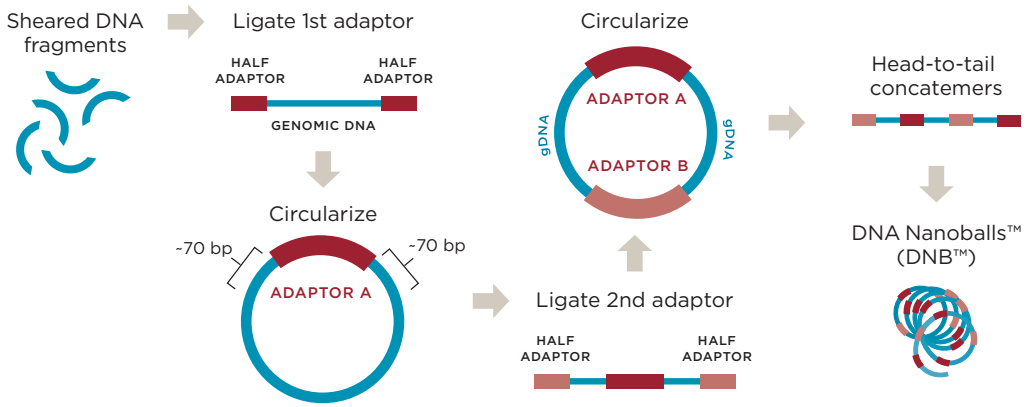
ASSEMBLY AND VARIANT CALLING

The Revolocity system can extract genomic DNA from blood or saliva samples, or use genomic DNA directly, as input for library construction.

The Revolocity Supersequencer is a high-capacity, fully integrated sample-to-variant calling system with the flexibility and scalability to meet changing human genome sequencing needs. From sample preparation, library construction, and sequencing to variant calling, the system automates and seamlessly integrates processes through an intuitive and user-friendly workflow management system. Preconfigured whole exome sequencing (WES) and whole genome sequencing (WGS) workflows make it easy to get started.

Innovations underlying the Revolocity library and sequencing chemistries and the WGS bioinformatics pipeline enable accurate small and large variant calls, from single nucleotide variants (SNVs), insertions, deletions, and small block substitutions to structural variants (SVs) and copy number variants (CNVs) across the genome. Data is provided in industry standard file formats such as BAM and VCF. Here, we describe the technological and algorithmic innovations in Revolocity that ensure excellent data quality.

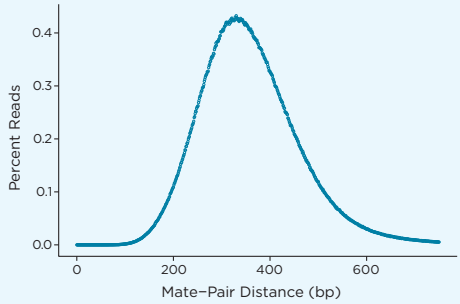
Proprietary library preparation—the first step to quality results



Library construction begins with DNA fragmentation, followed by the addition of two synthetic DNA sequences called “half adaptors,” one to each end of the sheared fragments. The molecules are then circularized through intramolecular ligation to form a complete adaptor. A series of proprietary methods shorten and linearize the DNA circles. A second set of half adaptors is added and the linear molecules are recircularized.

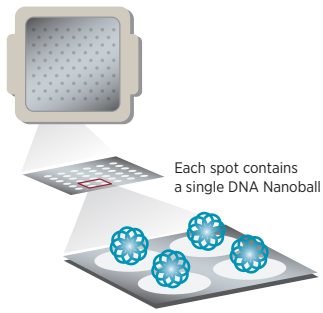
The resulting library is composed of DNA circles, which contain ~70 base pairs (bp) from each end of the original genomic DNA fragment, separated by the two adaptors. The genomic distance between the DNA fragments (referred to as “mate-pair distance”) is 250–500 bp. The tight distribution of the mate-pair distance has important implications for the quality of variant calls across the genome.

DISTRIBUTION OF MATE-PAIR DISTANCE



DNA Nanoballs™ (DNB™) optimize flow cell efficiency. The circular DNA molecules in the library are clonally amplified and modified to produce DNA Nanoballs (DNBs). DNBs are small particles with distinct size, density, and binding affinity properties, each containing more than 200 copies of the original template. When introduced to the patterned surface of a flow cell, the DNBs "self-assemble" into high-density DNB arrays, with >95% occupancy of flow cell spots occupied by a single DNB. The unique property of DNBs, combined with the design of the Revolocity flow cell ensures maximum flow cell utilization and system efficiency. With a single DNB per spot, and each DNB containing 200 copies of DNA, the result is brighter spots and high signal-to-noise ratio.

FLOW CELL

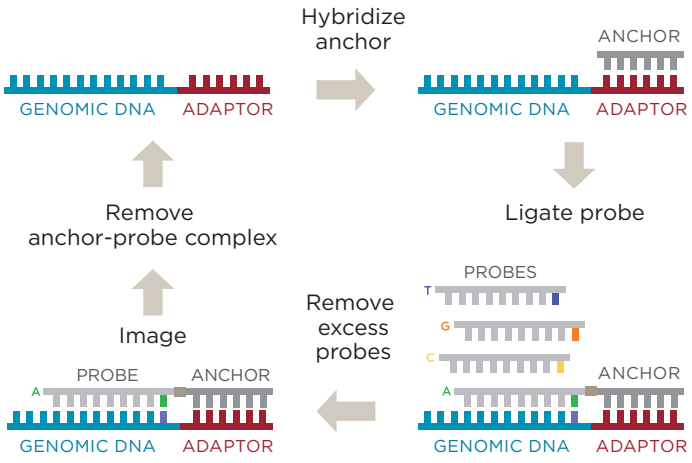


Combinatorial Probe-Anchor Ligation™ (cPAL™) chemistry

The Revolocity system employs Complete Genomics' well-established ligation-based cPAL sequencing chemistry[1] to generate a pair of 28 contiguous base pair (2 X 28) reads. The cPAL process comprises iterative cycles of anchor hybridization, probe ligation, imaging, and removal of the anchor-probe complex.

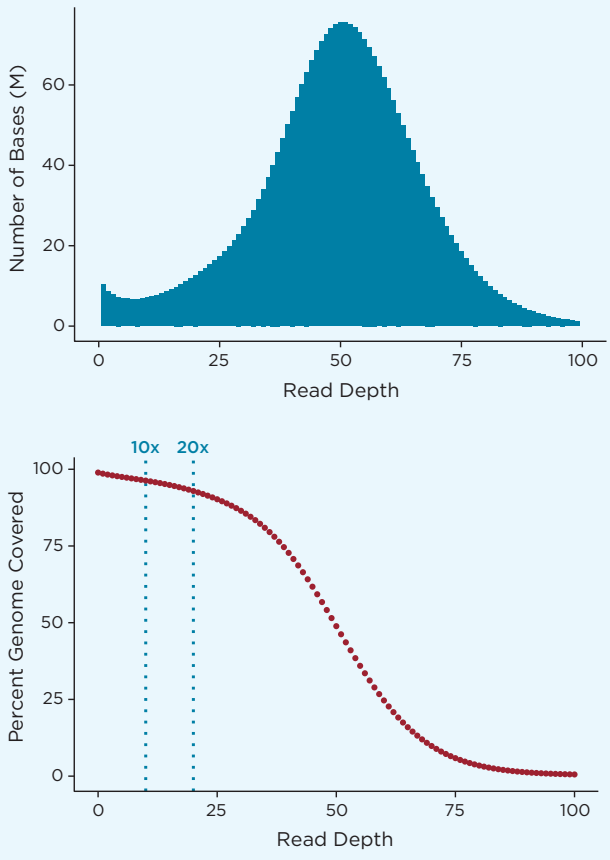
A sequencing cycle starts with the hybridization of an anchor sequence to one of the synthetic adaptors. To sequence a given position, a pool of probes is introduced with fluorescent labels associated with bases at that position. When the probe with the appropriate sequence hybridizes adjacent to the anchor, the anchor and the probe ligate to each other, thus stabilizing the interaction between the probe and the template DNA molecule. Unligated probes are washed away and the flow cell is imaged by simultaneous high-speed four-color detection. After a base is read, the entire anchor-probe complex is removed and the substrate is ready for another cycle of sequencing, using a different combination of probes and anchors to interrogate a different base position.

CPAL SEQUENCING TECHNOLOGY



cPAL advantages. A major advantage of cPAL over other sequencing technologies such as sequencing-by-synthesis is that a base-read cycle does not depend on the completeness of any of the previous cycles. If a base cycle fails, it does not prevent interpretation of the remaining bases for that DNB. This fault tolerance is not provided in other technologies, where a failed sequencing cycle impacts the rest of the read and may require sample resequencing.

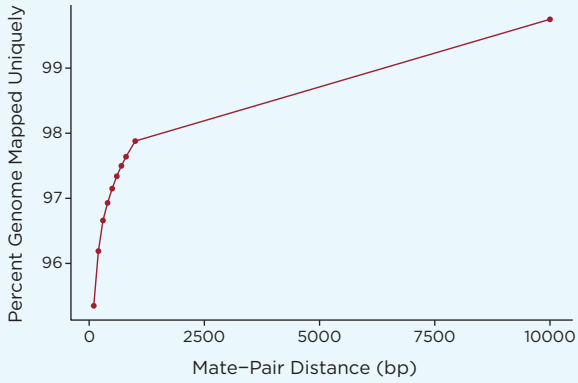
GENOME COVERAGE



cPAL provides highly uniform coverage with 97% and 94% of the genome covered at 10x and 20x depth of coverage, respectively.

Detecting variants across >97% of the human genome. cPAL sequencing chemistry generates 2 X 28 contiguous base pair reads with mate-pair distance in the range of 250–500 bp. With a mate-pair distance of 100 bp, 28-bp mate-pair reads can map uniquely to ~95% of the genome. Increasing the mate-pair distance to 400 bp allows ~97% of the genome to be mapped uniquely. By virtue of its mate-pair distance, the Revolocity system can call variants with high confidence in >97% of the genome and can even provide information about regions of the genome where there is insufficient evidence to make a reference or a variant call.

IMPACT OF MATE-PAIR DISTANCE ON MAPPABILITY

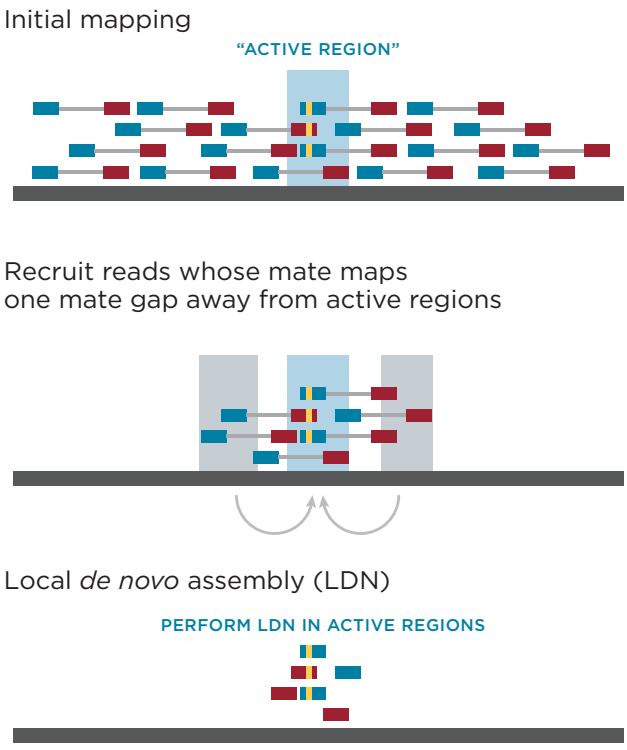


The Revolocity assembler: optimized for performance. The Revolocity assembler is a collection of multiple algorithms developed in conjunction with cPAL sequencing technology to generate highly accurate variant calls.

To identify small variants across the genome, reads are initially mapped to the reference genome in a process designed to identify genomic regions that deviate from the reference sequence. The mapping is tolerant of small deviations from the reference sequence, such as those caused by SNVs, read errors, or unread bases. Regions that are likely to differ from the reference genome, referred to as "active regions," are first identified, and reads that are likely to lie in these active regions are gathered from reads in which one arm maps approximately a mate-pair distance away from the active region. Next, local *de novo* assembly within active regions is performed by constructing and traversing a De Bruijn graph. The assembly software then optimizes the called sequence and uses evidential (Bayesian) reasoning to detect and score variants.

This approach was pioneered by Complete Genomics and results in highly accurate variant calls across the genome.[2] Because local *de novo* assembly is performed in active regions ~200 bp long, the Revolocity assembler provides accurate variant calls for SNVs as well as for insertions and deletions.

REVOLOCITY ASSEMBLER
SMALL VARIANT CALLING ALGORITHM



"No-calls." The Revolocity assembler identifies regions of the genome where sequence evidence is insufficient to make a reliable reference or variant call (i.e., "no-calls"). It provides an explicit distinction between homozygous reference calls and no-calls. Making no-calls provides significant advantages in genome comparison, for example by identifying false-positive *de novo* variants, when a region is variant in proband but no-called in parents. Other variant calling algorithms simply identify and report variant positions across the genome with the assumption that the remainder of the genome is homozygous reference. For positions where variants are not called, one must rely on surrogate measures such as depth of coverage and mapping scores to interpret whether nonvariant sites are homozygous reference or are simply not callable.

Reducing false-positive rate in small variant calls. The Revolocity assembler software includes a "correlation analysis" feature, where two active regions that are supported by the same reads are evaluated together. This enables the Revolocity assembler to identify and remove dubious variant calls produced by reads that map to multiple locations. The importance of correlation analysis is illustrated by an analysis performed to compare accuracy before and after applying the correlation filter. The results show a dramatic improvement in call quality—the correlation filter effectively identifies dubious variation calls that are likely artifacts of sequence similarity between different regions of the genome.

IMPACT OF CORRELATION ANALYSIS ON FALSE-POSITIVE RATE

Outcome	Without correlation analysis	With correlation analysis
SNP Sensitivity	98.5%	98.3%
SNP False-Positive Rate	12%	2.1%
Indel Sensitivity	83.4%	83.1%
Indel False-Positive Rate	26.3%	17.8%

Detecting CNVs and SVs. CNVs are a special case of large-scale variations in individual genomes relative to reference sequence that are manifested either as deletions of certain genomic regions or as the insertion of one or more copies of a region of the genome. While some CNV detection algorithms address CNV prediction using read depth, the Revolocity assembler uses both GC-corrected depth of coverage and continuous signal interpolation at every genomic position to predict CNVs down to 100 bp.

Structural variation information provided by the Revolocity assembler includes detected SV breakpoints and event interpretation such as gene fusions, inversions, and deletions. The Revolocity assembler searches for clusters of anomalous pairs of arm mappings and then performs a local *de novo* assembly. During local *de novo* assembly, evidence for the putative SV is assessed. Additionally, the SV junction is assembled to identify the breakpoints at base-pair resolution as well as any transition sequence that might exist between the two junction ends.

Summary

Designed for whole genome and whole exome sequencing, the Revolocity system has automated and integrated every step of the workflow and leveraged Complete Genomics' well-established cPAL chemistry and proprietary assembly algorithms to produce high-quality data and accurate variant calls. The result is a flexible, scalable system that can meet expanding sequencing needs now and in the future.

www.CompleteGenomics.com

TOLL-FREE IN THE US OR CANADA 1.855.CMPLETE (1.855.267.5383)

INTERNATIONAL 1.650.943.2800

salesinquiries@completegenomics.com

REFERENCES

1. Drmanac R et al. Science. 2010 Jan 1;327(5961):78-81. doi: 10.1126/science.1181498. Epub 2009 Nov 5.

2. Carnevali P et al. J Comput Biol. 2012 Mar;19(3):279-92. doi: 10.1089/cmb.2011.0201. Epub 2011 Dec 16.

Published in the USA, September 2015, BR-TOV-01

Complete GENOMICS™ A BGI Company