

Convert contigs into chromosomes to produce gold and platinum genomes

The Proximo™ Platform utilizes the ultra-long-range information provided by proximity ligation (Hi-C) to produce high-quality, highly contiguous genomes with full chromosomes and haplotype resolution.


Introduction

The genome holds the molecular blueprints for life. For humans, animals and plants alike, it encodes the secrets of adaptation, nutrition, health, disease and aging. As we are entering an era of personal genome sequencing to enable precision medicine, there is also an increased demand for high-quality plant and animal reference genomes to support agriculture, conservation, and research. In stark contrast to the hundreds of thousands of human genomes sequenced in the past few years¹, fewer than 0.1% (<15,000) of the 10 – 15 million eukaryotic species on earth have been completely or partially sequenced.²

Broad-based access to rapid, low-cost, next-generation sequencing (NGS) has stimulated many large-scale human, animal and plant sequencing projects. However, newly sequenced genomes are

typically fragmented, incomplete, interspersed with sequencing and annotation ambiguities and errors, and without critical haplotype information.

Phase Genomics' Proximo Platform enables the production of “gold” and “platinum” quality eukaryotic genomes. Proximo kits employ proximity ligation (also known as Hi-C) to measure the physical proximity between DNA sequences in the cell. The three-stage Proximo genome scaffolding algorithm uses scaffolding optimization to group and order sequencing data into chromosome-scale scaffolds. Assemblies that are of a higher quality and contiguity allow for improved annotation accuracy, thereby facilitating the functional and comparative genomic studies needed to answer important questions in medicine, agriculture, and conservation.



THE ERA OF
**PLATINUM
GENOMES**
HAS ARRIVED

Phased Chromosome-Scale
Genome Assemblies with Hi-C

- In October 2018, PacBio and Phase Genomics announced the highest-quality, most contiguous, individual human genome assembly to date—and the first chromosome-scale, diploid assembly accomplished with only two technologies (PacBio® sequencing and Proximo Hi-C).
- The sample was from a Puerto Rican female. The publicly available assembly (PacBio HG00733) represents the nearly complete DNA sequence from all 46 chromosomes inherited from both parents.
- Sequencing was performed on the PacBio Sequel® II System. Hi-C data was generated with the Proximo Hi-C Kit. Assembly, scaffolding, and haplotype phasing were performed using a combination of the FALCON-Unzip, FALCON-Phase™ and Proximo algorithms.³

Experimental Strategy: Proximity Ligation

Hi-C is one of a number of "chromosome conformation capture" (3C) methods originally designed to study the spatial organization of chromatin.^{4,5} This technology employs cost-effective, high-throughput, short-read sequencing to identify genomic loci that are in close proximity in three-dimensional space but may be megabases apart in the linear genome sequence (or even on different chromosomes). This powerful methodology has enabled dramatic improvements in genome assembly as well as variant and epigenetic analysis.⁶ In addition, it has unlocked many applications in metagenomics and microbiology.⁷

The principles of Hi-C and proximity-guided genome assembly (PGA) are outlined in Figure 1. Proximity ligation library preparation has been streamlined and optimized by Phase Genomics, and is available in the form of easy-to-use Proximo kits (optimized for different sample types), or as comprehensive services.

Computational Strategies

Proximo

De novo whole-genome sequencing of eukaryotic organisms is challenging due to large genome sizes, variable ploidy, extensive duplications, repetitive elements/regions, and areas of high GC or AT content. Draft genomes assembled from short reads are typically comprised of thousands of contigs, with as many gaps and errors. Without the long-range information provided by Hi-C, contigs are often misjoined, resulting in incorrect assemblies.

The Proximo computational tool was developed by Phase Genomics bioinformaticians to leverage proximity information in order to enable accurate, chromosome-scale genome construction for virtually any organism. Proximo utilizes Hi-C data to assign contigs to chromosome scaffolds, arrange contigs into a linear ordering, and then orient the contigs in such a way as to maximize the likelihood of having generated the observed Hi-C data (Figures 1 and 2).

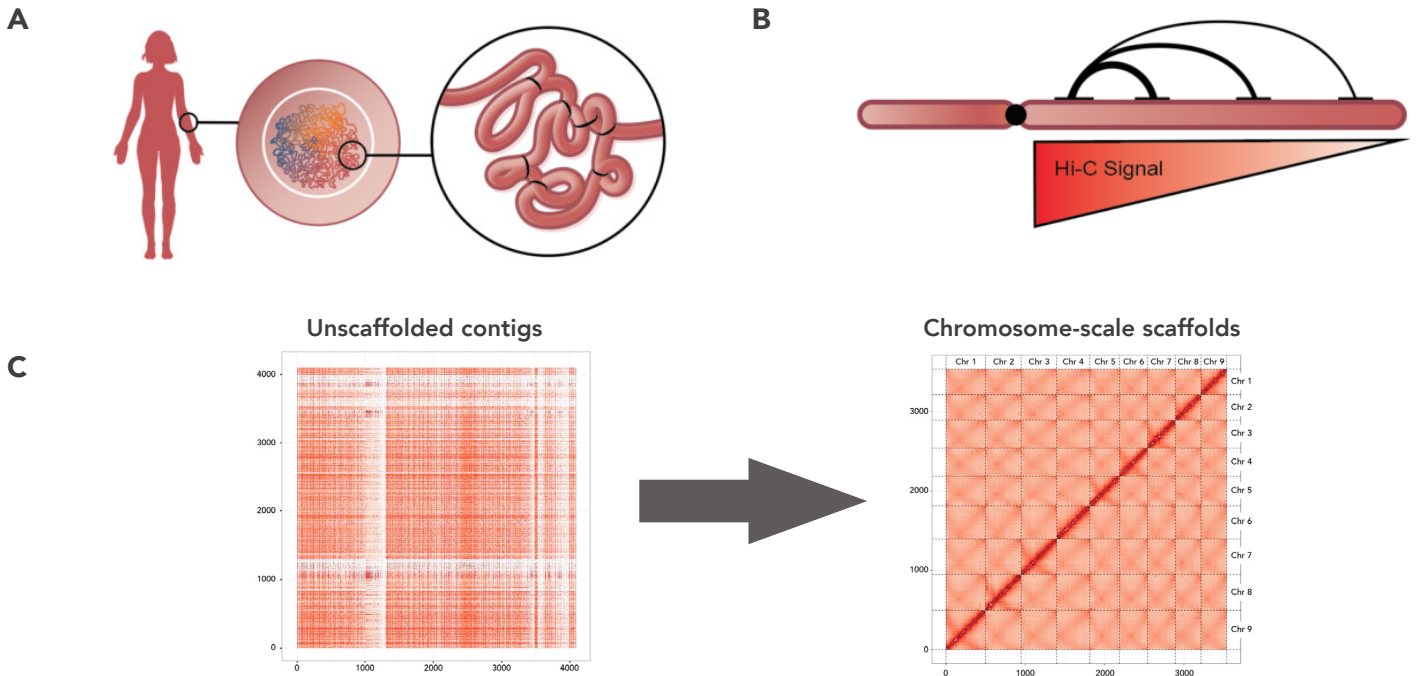


Figure 1. Overview of proximity ligation-based genome scaffolding. **A.** *In vivo* crosslinking traps both short- and long-range intracellular DNA contacts (depicted by black arcs). Crosslinked loci are fragmented and proximity-ligated, creating chimeric junctions originating from the same cell. These chimeric junctions are recovered, converted into an Illumina® sequencing library, and subjected to paired-end sequencing. **B.** The Hi-C signal increases as the genomic distance between any two loci across the genome decreases. **C.** The Proximo algorithm clusters scaffolded contigs (left) into chromosome groups. Contigs are ordered and oriented to produce a heat map (right). Each point on the heat map represents the depth of interaction between two genomic loci, captured in a chimeric junction. Loci that are closer to one another in two-dimensional space (on the genome sequence) are likely to generate more contact points (i.e. a darker color in the heat map). Using this principle, a whole genome assembly can be sorted into chromosome-scale scaffolds of any size, with a high degree of confidence.

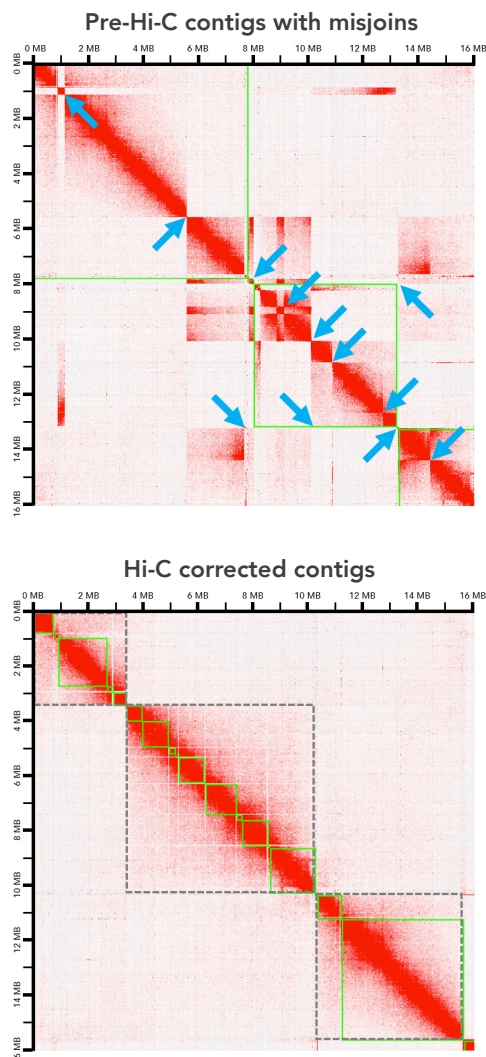


Figure 2. Hi-C enables high-quality assemblies by correcting misjoined contigs and generating chromosome-scale scaffolds. This example depicts Hi-C data mapped to a long-read eukaryotic genome assembly with multi-megabase-scale contigs. Before the application of Hi-C data (top panel), the assembly consisted of three large contigs (green blocks). When superimposed on the Hi-C heat map, numerous misjoins (blue arrows) and chimeric contigs are evident. The bottom panel shows the resulting scaffolds after misjoins were broken, and contigs were correctly oriented and scaffolded onto the three chromosomes (demarcated with dashed lines) using Proximo. Visualization adapted from Juicebox.⁸ Data courtesy of Dr. Kevin Solomon, Purdue University.

The core scaffolding algorithm is combined with an optimization process that performs tens to hundreds of thousands of iterations to find the scaffold solution most concordant with the data. Proximo is the only Hi-C scaffolding algorithm capable of directly consuming linkage maps or reference genomes, providing the ability to use more data as input to generate the best possible scaffolds.

Haplotype phasing with FALCON-Phase™

Homologous chromosomes are physically separate DNA molecules in the nucleus and, as such, form independent Hi-C profiles that can be used to identify which heterozygous sequences originated on the same chromosome homolog. FALCON-Phase,⁹ an algorithm co-developed by PacBio and Phase Genomics, examines contigs and haplotigs generated from long-read sequencing (e.g. the results of FALCON-Unzip¹⁰ or *purge_haplotigs*¹¹) in the context of Hi-C data, using a graph partitioning algorithm that detects likely phase switch errors and corrects them. This results in >96% contig phasing accuracy in known-truth, pedigree-based benchmarks (Figure 3).

FALCON-Phase can be used in conjunction with Proximo to extend phase blocks to chromosome scale, delivering two complete, truly phased sets of chromosomes for diploid organisms (both parental genomes, from a single analysis). FALCON-Phase is an open-source tool, available on GitHub (<https://github.com/phasegenomics/FALCON-Phase>), and as an analysis service from Phase Genomics.

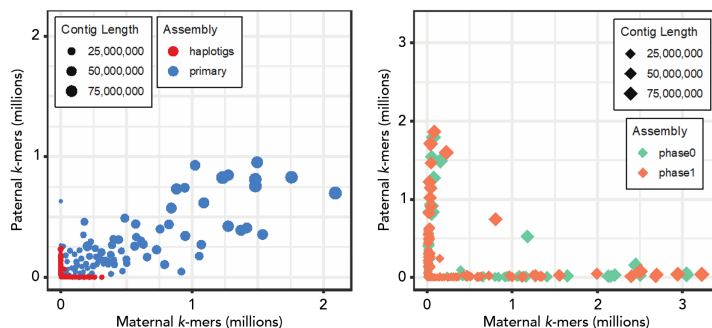


Figure 3. FALCON-Phase combines long-read and Hi-C data to generate accurate, fully-phased genome assemblies.

A trio-binned assembly was previously generated for a *Bos taurus* x *Bos taurus indicus* cross.¹² Chimeric contigs were broken by visualization of Hi-C read density with Juicebox. Hi-C read pairs were subsequently aligned to both the haplotig-containing phase blocks and the collapsed regions (portions of the primary contig without haplotigs), using BWA-MEM¹³. Both Hi-C read pairs were required to have a map quality >10, which yielded a haplotype-specific set of Hi-C reads. The counts of retained Hi-C read pairs mapping between phase blocks were used to generate a matrix, after which the phasing algorithm was applied. Graphs show the phasing accuracy of contigs before (left) and after (right) application of FALCON-Phase. Unphased, primary contigs (blue) are large, but contain a mixture of maternal and paternal markers. Haplotigs (red) are largely phased, but are significantly shorter. After phasing with FALCON-Phase, the phase0 and phase1 contigs are of similar length to the primary contigs and have significantly less mixing of parental markers.⁹

Applications

Gold and platinum quality genomes enable high-quality functional and comparative genomics studies

Annotation of (reference) genomes enables researchers to demarcate genes and regulatory sequences, map genetic variation, define genome structure, refine gene models and study gene function—in other words, to really understand the blueprints of the organisms that share our world. But insights can only be as good as the information from which it is derived.

The Proximo Platform has enabled the improvement of draft assemblies, and the generation of high-quality, chromosome-scale, haplotype-resolved reference genomes for humans, as well as scores of animals, plants and fungi (see Table 1 for examples). These gold and platinum genomes have contributed to advances in molecular and marker-assisted breeding, improved disease resistance and food security; and have supported breakthroughs in drug discovery and synthetic biology. In addition, they have enabled a better understanding of plant and animal evolution, adaptation and biodiversity.¹⁴

Table 1. Select eukaryotic genomes generated or improved with the Proximo Platform.

Organism	Genome Assembly Size	Final number of scaffolds	Scaffolded length (%)	Final N50 length	Reference
Human	2.74 Gb	23	98.02	125.7 Mb	1
Goat	2.62 Gb	31	98.74	91.7 Mb	2
Hummingbird	1.41 Gb	37	99.51	38.0 Mb	3
Clownfish	904 Mb	24	97.97	38.1 Mb	4
Firefly	473 Mb	10	94.64	49.2 Mb	5
Stickleback	446 Mb	21	97.52	20.6 Mb	6
Amaranth	400 Mb	16	98.09	24.1 Mb	7
Black raspberry	291 Mb	7	100	41.1 Mb	8
Honeybee	223 Mb	7	98.4	31.86 Mb	9
<i>Tolypocladium inflatum</i> (strain CBS714.70)	29.9 Mb	7	99.6	5.07 Mb	10

¹Burton et al. *Nat. Biotechnol.* 2013; 31: 1119. ²Bickhart et al. *Nat. Genet.* 2017; 49: 643. ³Pennisi, *Science* 2017; 357: 10. ⁴Lehmann et al. *BioRxiv* March 2018. ⁵Fallon et al. *BioRxiv* December 2017. ⁶Peichel et al. *J. Hered.* 2017; 108: 693. ⁷Lightfoot et al. *BMC Biol.* 2017; 15: 74. ⁸VanBuren et al. *GigaScience* 2018; 7: gjy094. ⁹Wallberg et al. *BMC Genomics* 2019; 20: 275. ¹⁰Olarte et al. *BMC Genomics.* 2019; 20: 120.

For a more complete list, visit www.phasegenomics.com/publications/#papers

Genomic contiguity information enables the detection of structural variants across the entire genome

Structural variants (SVs) are typically regarded as genomic alterations that involve DNA segments larger than 1 kb; and include insertions, deletions, inversions, duplications, translocations and copy-number variants (CNVs). Structural variation is associated with millions of bases of heterogeneity within every genome, and is now recognized as one of the primary contributors to the genetic and phenotypic variation that underlies human diversity and disease susceptibility.¹⁵

Because of their nature, SVs are difficult to study with confidence using short-read sequencing technologies. The Proximo Platform utilizes Hi-C to scan the structure of a genome, capturing both short-range and long-range genomic contiguity. When a reference genome is available, the technology enables the identification of structural differences between an individual and the reference; as SVs change the frequency with which different loci in the genome interact. ProximoSV comprises a suite of algorithms for the identification of structural and copy number variation in organisms with reference genomes (Figure 4). By combining several approaches that interrogate reference-aligned Hi-C data for aberrations, ProximoSV can be used to detect and classify these events, yielding a comprehensive view of large structural and copy number variants in a genome.

Because Hi-C allows for the genome-wide examination of all interacting loci, the Proximo Platform may also be used for the identification and high-resolution characterization of the structural elements of chromatin architecture. These elements range from loops between loci that are <1 Mb apart, to hubs of inter-chromosomal contacts and chromosomal compartments (Figure 5). Within this compartmentalization, topologically associating domains (TADs) are chromatin regions that interact more frequently within themselves than between each other. TADs can be highly conserved across species and cell types. The demarcation of TAD boundaries (TAD calling) facilitates the functional annotation of genes, and may contribute to the understanding of cell differentiation, development, and adaptation; as well as genetic diseases and cancer.¹⁶

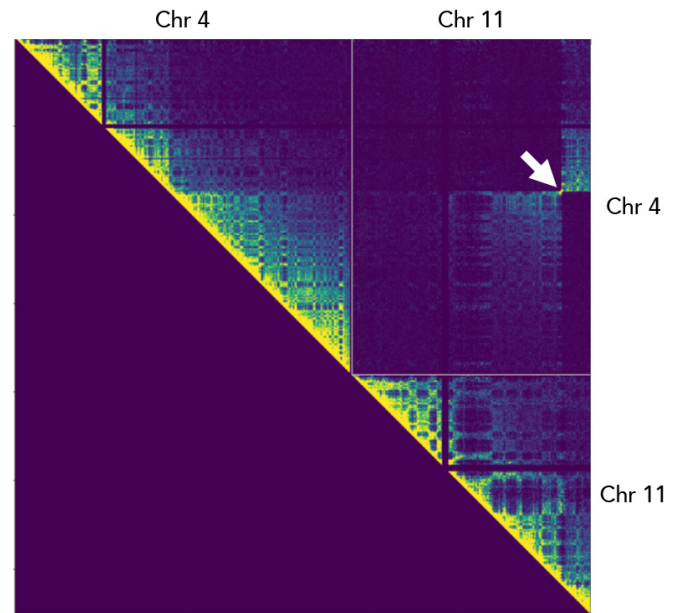


Figure 4. ProximoSV enables accurate calling of large structural variants from short-read sequencing data. Hi-C interaction heat map representing select portions of chromosomes 4 and 11 of a human genome. Yellow shading represents the strength of proximity interactions. A decrease in color intensity (green to blue to purple) is, on average, proportional with increasing physical distance. The area of high-intensity color in the top right-hand quadrant (white arrow) indicates a putative translocation between chromosomes 4 and 11.

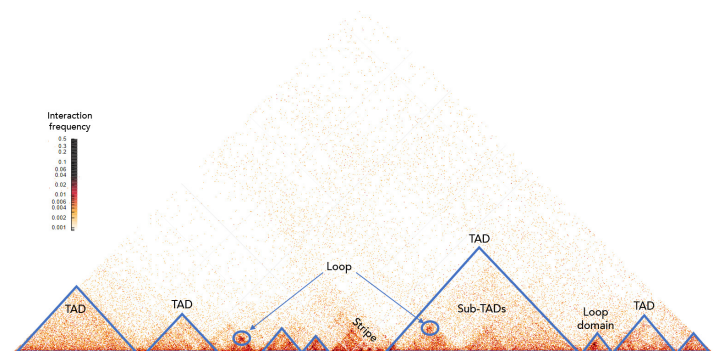


Figure 5. Chromatin structural elements observed in a Hi-C map. This heat map corresponds to a 4 Mb-window of human chromosome 1. Each pixel represents an interaction between two genomic loci. Color intensity represents the frequency of the interactions between each pair of loci; the darker the color, the more frequent the interactions. Triangular areas represent the regions in the genome where chromatin interactions occur. Most of the TADs are designated with blue triangles. Some TADs contain smaller sub-TADs. A darker dot on the apex of a triangle typically corresponds to a chromatin loop. A stripe (or track) represents dynamic structures that can be formed upon the release of individual CCCTC-binding factor molecules from chromatin.¹⁷ Visualization adapted from HiGlass.¹⁸

Summary

The Proximo Genome Scaffolding platform utilizes proximity ligation (Hi-C) data to group and order sequences into high-quality, chromosome-scale scaffolds. Hi-C data and Proximo may also be integrated with other computational approaches (e.g. FALCON-Phase™), to enable the production of haplotype-resolved gold and platinum quality eukaryotic genomes. Higher quality assemblies allow for improved annotation accuracy, thereby facilitating functional and comparative genomics.

Key features and benefits of the platform are summarized below:

- Proximo Library Prep Kits provide a streamlined Hi-C protocol optimized for different sample types (human, plant, animal, fungal or microbial) that does not require the extraction of high-molecular weight (HMW) DNA.
- Cost-effective short-read sequencing data may be combined with draft assemblies obtained from short- or long-read sequencing.
- Available as user-friendly kits with Proximo Genome Scaffolding Analysis service, or as a comprehensive sample-to-analysis service.



Learn more about the Proximo Platform at phasegenomics.com/products/proximo/

References

1. <https://www.lesswrong.com/posts/geE9t5Dm9iq6Y7nQ4/progress-review-genome-sequencing-june-2019>
2. Lewin HA, et al. Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci USA* 2018; 115(17): 4325-4333. doi: 10.1073/pnas.1720115115.
3. https://www.pacb.com/press_releases/pacific-biosciences-releases-highest-quality-most-contiguous-individual-human-genome-assembly-to-date/
4. Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009; 326(5950): 289-293. doi: 10.1126/science.1181369.
5. Van Berkum NL, et al. Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* 2010; 39: e1869. doi: 10.3791/1869.
6. <http://phasegenomics.com/applications/human-genomics-epigenomics/>
7. <http://phasegenomics.com/applications/metagenomics-microbiology/>
8. Durand NC, et al. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* 2016.; 3(1): 99-101. doi: 10.1016/j.cels.2015.07.012.
9. Kronenberg ZN, et al. Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nat. Commun.* 2020. Accepted.
10. Chin SC, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 2016;13(12):1050-1054. doi: 10.1038/nmeth.4035.
11. Roach MJ, et al. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* 2018; 19: 460. doi: 10.1186/s12859-018-2485-7.
12. Koren S, et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* 2018; 36: 1174–1182. doi: 10.1038/nbt.4277
13. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN] (2013).
14. <https://phasegenomics.com/applications/plant-animal-genomics/functional-comparative-genomics/>
15. Spielmann M, et al. Structural variation in the 3D genome. *Nat. Rev. Genet.* 2018; 19: 453–46. doi: 10.1038/s41576-018-0007-0.
16. Zuffrey M, et al. Comparison of computational methods for the identification of topologically associating domains. *Genome Biol.* 2018; 19: 217. doi: 10.1186/s13059-018-1596-9.
17. Mota-Gómez I and Lupiáñez DG. A (3D-Nuclear) space odyssey: making sense of Hi-C maps. *Genes* 2019; 10: 415. doi: 10.3390/genes10060415.
18. Kerpedjiev P, et al. HiGlass: Web-based visual comparison and exploration of genome interaction maps. *Genome Biol.* 2018; 19: 125. doi: 10.1186/s13059-018-1486-1.



info@phasegenomics.com
www.phasegenomics.com

Phone: 1-833-742-7436
Twitter: @PhaseGenomics

Unless otherwise stated, data on file.

For Research Use Only. Not for use in diagnostic procedures.

PROXIMO and FALCON-PHASE are trademarks of Phase Genomics, Inc. All other product names and trademarks are the property of their respective owners.

© 2020 Phase Genomics, Inc. All rights reserved.

AN002-091020