Genomics, Proteomics & Bioinformatics

Long-Range NGS Linked-Reads and Applications of Hi-C, 10x, Haplotagging and TELL-Seq Platforms

--Manuscript Draft--

Manuscript Number:	GPB-D-22-00474
Article Type:	Review Article
Keywords:	Long-range NGS reads, Hi-C, 10x, haplotagging, TELL-Seq, genome assembly and quality assessment
Corresponding Author:	Zemin Ning, PhD The Wellcome Sanger Institute ambridge, UNITED KINGDOM
First Author:	Zemin Ning, PhD
Order of Authors:	Zemin Ning, PhD
Abstract:	Long-range reads grant insight into additional genetic information, from the original DNA samples, far beyond what can be accessed by short reads, or even modern long-read technology. Several new sequencing technologies have become available for long-range “linked reads” with high-throughput and high-resolution genome analysis. These long-range technologies are rapidly advancing the fields of genome assembly, genome scaffolding, and allowing more comprehensive variant identification. In this article, we focus on four major long-range sequencing technologies: Hi-C, 10x Genomics Linked-Reads, Haplotagging and TELL-Seq. We detail the mechanisms and data products of these four platforms, introduce their most important applications, evaluate the quality of sequencing data from different platforms, and discuss the currently available bioinformatics tools. We hope this work will benefit the selection of appropriate long-range technology for specific biological studies.
Suggested Reviewers:	Yingguang Frank Chan, PhD Professor, Max Planck Institute for Biology Tübingen frank.chan@tue.mpg.de Dr Chan invited Haplotagging and is a leading expert in long-range data analysis.
	Joana Isable Meier, PhD Group Leader, University of Cambridge and Sanger jm2276@cam.ac.uk Haplotagging expert
	Chenxi Zhou, PhD Postdoc, University of Cambridge cz370@cam.ac.uk Dr Zhou developed YaHS, a popular genome scaffolding tool which has been widely used.
	Arang Rhie, PhD Staff scientist, NIH/NHGRI arang.rhir@nih.gov Dr Rhie developed a number of widely used informatics tools, including Merqury for assembly and data QC.
Opposed Reviewers:	

Dear Editor,

We are pleased to submit our manuscript entitled "Long-Range NGS Linked-Reads and Applications of Hi-C, 10x, Haplotagging and TELL-Seq Platforms" by Jiang *et al*., for publication as a review/research article in Genomics, Proteomics & Bioinformatics.

With rapid development in sequencing technologies, long-range data types and their applications have gained substantial momentum in the genomic community. Long-range reads grant insight into additional genetic information, from the original DNA samples, far beyond what can be accessed by short reads, or even modern long-read technology. The roles of these datasets in genome scaffolding, consensus base polishing, phasing as well as structural variation detection are simply unreplaceable, without which many high-profile projects such as VGP, Darwin Tree of Life and Earth BioGenome won't be able to achieve the targeted objectives. Currently, there is not a single published paper which discusses the common features, data evaluations and applications of long-range data, when looking at literature. We hope this paper fills this important gap and is of great interest in both genomics and bioinformatics.

This is an invited review paper and we haven't submitted this work to other journals. All the authors have seen and approved the submitted version of the manuscript. Please feel free to contact me, should you have any questions.

Dr Zemin Ning

Senior Scientific Manager
High Performance Algorithms
The Wellcome Sanger Institute
Wellcome Genome Campus
Hinxton, Cambridge CB10 1SA
UK
Tel: (44) 1223 494705

E-mail: zn1@sanger.ac.uk

Dear Editor,

We would like to suggest these outstanding researchers with expertise in individual areas as referees for this manuscript:

1. Haplotagging
   Professor Yingguang Frank Chan
   Max Planck Institute in Tuebingen
   frank.chan@tue.mpg.de

2. Haplotagging:
   Dr Joana Isabel Meier*
   University of Cambridge
   jm2276@cam.ac.uk

3. Hi-C and genome scaffolding
   Dr Chenxi Zhou
   University of Cambridge
   cz370@cam.ac.uk

4. Hi-C and 10x
   Dr Arang Rhie
   Genome Informatics Section at NIH/NHGRI,
   arang.rhir@nih.gov

   *Dr Meier also has a group at Sanger now with Tree of Life Project. But work wise, I have no overlaps with her. Haplotagging is new technology and there are not so many experts in the community right now. Given this is a review paper, I think it is legitimate to invite referees who are from the same institute.

This is an invited review paper and we haven't submitted this work to other journals. All the authors have seen and approved the submitted version of the manuscript. Please feel free to contact me, should you have any questions.

Dr Zemin Ning

Senior Scientific Manager
High Performance Algorithms

The Wellcome Sanger Institute
Wellcome Genome Campus
Hinxton, Cambridge CB10 1SA
UK
Tel: (44) 1223 494705

E-mail: zn1@sanger.ac.uk

Long-Range NGS Linked-Reads and Applications of Hi-C, 10x, Haplotagging and TELL-Seq Platforms

Libo Jiang[1], Michael A. Quail[2], Jack Fraser-Govi1[2], Haipeng Wang[1], Xuequn Shi[3], Karen Oliver[2], Esther Mellado Gomez[2], Fengtang Yang[1] and Zemin Ning[2+]

[1]School of Biological Sciences, Shandong University of Technology, Zibo, 255049, Shandong, China.
[2]The Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK
[3]College of Food Science and Technology, Hainan University, Hainan, 570228, China

[+]Corresponding author. E-mail: zn1@sanger.ac.uk (Ning Z).

Abstract

Long-range reads grant insight into additional genetic information, from the original DNA samples, far beyond what can be accessed by short reads, or even modern long-read technology. Several new sequencing technologies have become available for long-range "linked reads" with high-throughput and high-resolution genome analysis. These long-range technologies are rapidly advancing the fields of genome assembly, genome scaffolding, and allowing more comprehensive variant identification. In this article, we focus on four major long-range sequencing technologies: Hi-C, 10x Genomics Linked-Reads, Haplotagging and TELL-Seq. We detail the mechanisms and data products of these four platforms, introduce their most important applications, evaluate the quality of sequencing data from different platforms, and discuss the currently available bioinformatics tools. We hope this work will benefit the selection of appropriate long-range technology for specific biological studies.

Keywords

Long-range NGS reads, Hi-C, 10x, haplotagging, TELL-Seq, genome assembly and quality assessment

Introduction

Next-Generation Sequencing (NGS) technologies have revolutionized the field of genomics and genetics, providing low-cost and high-throughput data at an unprecedented scale. However, most NGS technologies make an underlying assumption that all relevant genetic information can be reconstructed from the smaller fragments that make up both short (100-250bp) and long (>10,000bp) reads. Such reads are 'short range' or 'local', because they contain only information about the genetic sequences of the reads, in contrast to 'long-range', 'non-local' or 'linked' reads, which retain additional contextual information regarding the origin of the read within the complex, 3-dimensional physical structure of the DNA within and between chromosomes.

We emphasise that, despite the similar terminology, long-range reads are conceptually distinct from long reads. Although the size of long reads provides a large quantity of information, it is inherently local, relating only to the sequence without containing additional information about the origin of the fragment. In contrast, long-range reads provide additional non-local information, and can take the forms of both short and long reads, though in practice most long-range technologies currently use short reads. To avoid confusion, we suggest deviating from the literature standard and instead defining 'long-range' reads as either non-local reads, or linked reads.

Without the additional context of non-local information, for example, it remains challenging to reliably identify structural variation (SV) with short reads. Although short reads can identify SVs to base-pair resolution, utilizing only short-range information suffers from a higher false discovery rate than long reads [1]. It is also difficult to phase many millions of short reads to a haplotype-resolved genome, particularly for highly repetitive sequences, complex heterozygosity, and large polyploid genomes [2, 3].

Local long reads can sidestep many of the issues associated with local short reads although they contain only local information, because the large size of the reads makes it much easier to uniquely localize them within the genome [4, 5] [6,7]. Currently, two major long-read technologies: Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing and Oxford Nanopore Technologies (ONT) nanopore sequencing are used for long-read genome analysis [4]. However, these

long-read methods have two drawbacks: (i) higher costs and lower throughput and (ii) higher DNA input requirements compared to short-read sequencing.

Whilst long-range information can be used in isoltation for *de novo* assembly, long-range reads have already proven vastly more powerful since the large, contiguous reads make referenceless assembly much easier. However, the throughput and cost issues associated with long reads mean that using them as the sole means of long-range information in the *de novo* assembly of tens of thousands of genomes will likely be prohibitive. This is especially true if chromosome-level assemblies are desired, since long reads are still much smaller than chromosomes and hence do not carry chromosome-scale context. Therefore, a cheaper method for inferring long-range information is needed.

Several mechanisms for storing long-range information within short reads and hence the necessary context to reconstruct a single long molecule of DNA have been developed, including the "Pair-Linked Reads" (PLR) chromosome conformation capture-based Hi-C and "Chain-Linked Reads" (CLR) technologies [8]. 10x Genomics provides perhaps the best known chained read strategy, which can generate long-range information from standard approaches based on short reads [9]. In recent years, 10x Genomics Chain-Linked Read technology has been widely used, but a variety of other barcode-based methods such as TruSeq, BGI's Long Fragment Reads, TELL-Seq, LoopSeq and haplotagging have been developed with ultralow DNA input, high per-base resolution, and low costs [10, 11, 12, 13].

Here, we focus on the four major long-range reads sequencing technologies, i.e., Hi-C, 10x Genomics Linked-Reads, haplotagging and TELL-Seq. Firstly, we detail the protocols and mechanisms of the four platforms' function. Secondly, we propose some criteria to evaluate the quality of sequencing data on different platforms, and apply these criteria to discuss the characteristics of datasets either downloaded from public resources or generated by us. Thirdly, we review the practical applications of these technologies in efforts such as genome scaffolding, de novo assemblies and variation screens. Finally, we provide a list of software tools which are commonly used for genome analysis with long-range reads and discuss their strengths and weaknesses.

Platforms

In this section, we briefly detail the four platforms of interest: Hi-C, 10x, Haplotgging and TELL-Seq, focusing on the protocols used to generate the long-range data, and how such long-range information is manifested in the data products. Across these four platforms, long-range, non-local information is stored in 'linked reads' in one of two ways: either in Pair-Linked Reads, in which two reads are coupled together to indicate a relationship between them, or Chain-Linked Reads, in which reads are tagged or labelled in some ways to indicate their origins. This qualitatively changes the non-local information provided by the platform, and hence informs which platform is suitable for a given application.

Hi-C

Hi-C is a Pair-Linked Read technology, and the culmination of several generations of Chromosome Conformation Capture technologies, which uses PLR to probe chromosome conformation – the spatial organization of chromatin within a cell – at a genome-wide scale [14], granting access to 3D proximity information within the nucleus. Since chromatin is a complex 3-dimensional structure, this information allows researchers to detect long-range interactions between segments within a chromosome or between different chromosomes. Since homologous chromosomes each tend to occupy distinct territories in nuclei [15], this feature enables the use of Hi-C data to improve *de novo* assembly, and phase heterozygous genome variants onto haplotypes,

Hi-C technology follows this protocol (see graphical summary in Fig.1A):

- The nuclear chromatin is crosslinked using formaldehyde. By design, these crosslinks occur preferentially between strands that are close together in 3D space.
- Crosslinked chromatin is solubilized and fragmented with a restriction enzyme
- The crosslinked segment ends are repaired by filling in with biotin-labeled nucleotides.
- DNA ligase is used to cyclize the blunt-end components, the proteins that bind the DNA fragments are degraded, and then the circular crosslinked fragments are

randomly broken again using sonication or other methods.

- DNA is purified and sheared. The biotin-labeled DNAs are captured with streptavidin-conjugated beads and amplified before sequencing.

The final result of this protocol is a large number of "deliberate chimeric", paired short reads, with each end of the read originating from one of the crosslinked strands, which are potentially very far away from each other in the linear genome – and even on different chromosomes entirely. The generated paired reads are then mapped to a contig assembly of the genome and used to create a high-resolution interaction map within and between chromosomes: regions where larger numbers reads are found to have been crosslinked are then inferred to be regions of close contact between the DNA strands.

10x Genomics linked-reads

10x Genomics Linked-Reads (henceforth simply '10x') are a product formerly provided by 10x Genomics. In 10x sequencing, long-range information is retained by combining 3' barcoding with standard short-read sequencing [16], producing short Chain-Linked Reads with a 'memory' of the larger scale locality where they were derived from, and hence making it easier to assemble the resulting dataset. The resulting reads can improve the quality of genome assembly by expanding the range of linking information along the chromosome to define haplotypes. The 10x protocol (demonstrated in Fig.1B) is as follows:

- First, high molecular weight (HMW) DNA is prepared and sheared into long DNA fragments (ideally > 100kb)
- At two microfluidic junctions, tHMW-DNA is then combined with an oil-surfactant solution, enzymes and gel beads loaded with random primers and barcode sequences to produce "Gel Beads in Emulsion" (GEMs). Each GEM captures around 10 HMW-DNA fragments.
- The GEMs are then isolated in partitions and the beads dissolved, releasing the barcodes and primers uniquely to the HMW-DNA fragments in that partition.
- Each partition is then sheared, extended with both the barcodes and the primers, and then amplified and sequenced – in this case, by Illumina paired-end sequencing.

The end result of this process is a number of short reads preappended by a unique barcode identifying the GEM bead they originated from: all reads sharing a common barcode are called ‘Linked-Reads’ (which we distinguish as being distinct from the general term linked reads: under our terminology these are Chain-Linked Reads, a specific form of linked reads). The key statistic is that, since each GEM captures so few HMW-DNA fragments, the odds that a second fragment which shares the same barcode also originates from nearby in the genome is very small (see Supplementary Figure 2), and so the barcode acts to (nearly) uniquely group sets of reads together as being spatially co-located. This, for example, makes it much easier to phase short reads as the entire barcoded molecule must be simultaneously phased.

Although 10x sequencing can reconstruct multi-megabase phase blocks by assembling short reads with barcode information, it still has some drawbacks, such as relatively high costs in library preparation, and that the 10x platform performs counterintuitively when faced with smaller genomes, showing a marked performance degredation. This is because the partitions get saturated by the smaller genome size, and the statistics begin to favour ‘barcode collisions’ much more frequently. The 10x platform is optimized for the human genome size, and modifications such as smaller DNA samples would need to be made for non-human cases.

Most significantly, however, this product was been withdrawn and discontinued in 2020 [17]. However, we include this platform in our analysis for continuity with previous benchmarking and comparison efforts, and since future 10x Genomics products may be comparable to this previous iteration.

Haplotagging

Several other technologies have been developed to provide an alternative form of CLR in the absence of 10x. Haplotagging, as a simple and relatively low-cost Chain-Linked Read sequencing technique was developed by Meier et al. [13]. This technique allows high-throughput sequencing without losing haplotype information while maintaining the power, accuracy, and scalability of standard Illumina sequencing.

Haplotagging, like TELL-Seq mentioned below, is a transposon bead-based technology that employs transposomes containing bead-specific barcoded adaptors. These technologies utilise the tendency of segments of HMW-DNA to wrap around

microbeads, providing many points of contact between the bead and the DNA. The full protocol is as follows (Fig.1C):

- As in the 10x protocol, HMW-gDNA (ideally >100kb) is prepared.
- The HMW-gDNA is mixed with the barcoded beads. Each bead carries a standard Illumina Nextera Tn5 transposon adaptor, augmented with one of 85 million barcodes, and each bead captures only a single DNA fragment
- Transposition transfers the barcoded adaptors into the long DNA fragments, before PCR amplification to generate a sequencing-ready library.
- Finally, the libraries were sequenced using an Illumina platform.

The result is that the initial HMW-DNA fragments are broken into smaller units, each containing a unique barcode, that can be sequenced on short read sequencers. Subsequently, all the reads originating from the same HMW-DNA fragment can grouped by their barcode, and hence correctly mapped to the same fragment.
The key difference between haplotagging technology and 10x is that DNA molecules tend to interact only with a single bead, instead of the approximately 10 (for humans) fragments-per-bead that 10x relied on. In addition, each bead is tagged with four barcode fragments that are distributed in the standard i5/7 index positions of the Illumina Nextera adaptor design. Thus, library preparation and barcoding are performed simultenously within the same tube, making the process cheap and easy to produce using standard molecular biology equipment. The data output is very similar to that of the 10x platform: a series of short reads preappended by a barcode, indicating which reads originated from a similar vicinity. However, as mentioned above, the process is much cheaper (the original work claimed a 99% cost reduction); and since the fragment/bead interaction is close to 1:1, instead of approximately 10:1, each fragment is genuinely uniquely barcoded, resulting in fewer barcode collisions, as demonstrated in Supplementary Figure 2. In addition, the 4-fragment nature of the barcode is designed to allow for error-correction in the barcode reads, allowing for more robust identification of the barcode. However, the fragments are prone to display PCR duplication errors [18] and the product is not yet at the stage of commercial deployment.

TELL-Seq

TELL-Seq[1] is another CLR sequencing technology which functions very similarly to the Haplotagging platform but is currently commercially available through Sage Science. The TELL-seq technology workflow is as follows (Fig.1D):

- Genomic DNA (0.1-5ng), the barcoded TELL beads (3-10 million) and transpososomes are mixed in a PCR tube.
- The transpososomes and DNA segments interact to form a strand transfer complexes (STCs), which is connected with the barcode sequence on the TELL bead surface.
- The transposase is removed, the DNA fragment is cut into two parts in the STC, and the beads removed, leaving a DNA fragment, connected to a transposon, which is in turn connected to a barcode.
- The barcoded DNA molecules are amplified with P5 and P7 adaptors before illumina sequencing.

The library preparation for TELL-Seq differs from haplotagging in mostly minor ways, with the primary distinction being in the form that the barcode takes: TELL-Seq uses a simpler 18-base barcode, rather than the 4x6 method of Haplotagging. This allows for a larger number of unique barcodes – and hence reduced likelihood of a collision - but lacks the error-correction feature.

Data features and quality assessment

Before we discuss data applications, we first introduce metrics on quality assessment and then use the metrics to evaluate datasets sequenced for this study. Our focus will be on Hi-C, 10x and Haplotagging, which are currently or previously available in the market.

[1] We note that the acronym TELL-Seq (Transposase Enzyme Linked Long-Read Sequencing) falls afoul of the terminology confusion referenced earlier. Under the terminology we have enforced, the linked reads produced are *long range*, but they are not *long reads*.

Data Metrics

In order to provide a robust analysis of the relative performance of the platforms, we must first derive numerical metrics by which to judge them. Since the Pair-Linked Platforms platforms differ significantly in the mode of operation from the Chain-Linked Reads, the metrics used will be slightly different, but our design aims to enable as valid a comparison as possible.

Metric 1: Association

Association is the ability for long range information to be communicated by the platform, or equivalently, the amount of non-local information contained within a read. Datasets with a higher association contain more and longer-range information than those with a lower association. In the context of trying to use long-range information as an assembly tool, a stronger association is preferable.

For the PLR platforms, association is measured by the distribution of Link-Separation Distance, the distance on the linear genome between the two ends of paired reads which have been linked together. If the first end of the pair aligns to a location i, and the second end to j, then the genomic distance is $|i-j|$. If large values of $|i-j|$ are found to occur more often, then the dataset has a stronger association. Whilst we should therefore favour platforms which have a higher proportion of reads with large $|i-j|$, we note also that there is an expected pattern at higher distances: if the linkage probability is inversely proportional to some power (b) of the physical distance between the reads, and at large linear distances genomic distance and physical distance are approximately the same, then we expect the frequency to fall as

$$p(\text{link } i,j) \propto \frac{1}{|i-j|^b} \quad \rightarrow \quad f(|i-j|) \approx A|i-j|^{-b}$$

Where A is an arbitrary scaling parameter. On a log-log scale, this manifests as a linear relationship between the separation distance, and the observed frequency.

Deviations from this pattern indicate problems with the library preparation and can result in the failure of any statistical inference based on the dataset. We should therefore prefer datasets which i) exhibit a power-law relationship in frequency at high separation distances and ii) Have a smaller exponent, resulting in a longer tail, and hence more long-range information.

In the case of the Chain-Linked Read platforms, the long-range information is conveyed by labelling reads as originating from a larger molecule via a tag shared by all fragments of that molecule. The association should therefore be measured by the size of the molecules from which the labelled reads are drawn.

It is clear that having a larger molecule is generally better: each barcode delinieates a larger spatial region, so the information is longer-range. There is, however, an upper limit at which point increasing the molecule size gives decreasing returns: for example, if the molecules were chromosome scale, then the barcoding would simply inform us which chromosome the read is from: useful, but not beneficial for assembling the reads within a given chromosome. Of critical concern, however, is that increasing the molecule size increases the chances of barcode collisions, behaviour demonstrated in Supplementary Figure 2. Generally, the size at which collision rates become untenable is significantly below the genome size, and hence should be treated as the limiting factor on the molecule size. We should therefore favour platforms which generate larger molecule lengths, but which still have a small collision rate.

Metric 2: Accessibility

Accessibility is the fraction of the data which is unique, unambiguous and useable. Datasets which have a low accessibility may still contain useful scientific data, but much more data would be required to achieve the same level of significance. We should therefore prefer platforms which produce highly accessible data. For example,

both CLR and PLR suffer from potential PCR duplication – the overamplification of some portions of the genome through the library preparation process. A high PCR duplication rate is indicative of a poor accessibility, and vice versa. Complex factors underlying the library preparation can also lead to reads which cannot be mapped to the reference genome (and the rate of unmapped reads is noteably higher in Long Range platforms than normal Illumina short reads), or which contain no linking information ('singletons'). Such 'unmapped' reads contain no useful information, and so they too should be excluded from further analysis.

In addition, PLR explicitly allows inter-chromosomal interactions to be mapped. Whilst this is useful in general in 3D genomics, for the purposes of the applications discussed in section 0 this represents unusable data, as assembly should occur on a per-chromosome basis. In order to have the maximum amount of usable information, we should therefore prefer the platforms which have a smaller number of linkages between chromosomes: a smaller translocation rate.

Assuming that other sources are negligible (or, equal between platforms), the total accessibility of the dataset can therefore be computed from the PCR duplication rate D, the translocation rate T and the unmapped rate, U:

$$A = 1 - D - T - U$$

A higher value of A indicates a dataset which contains more useful information.

Metric 3: Evenness

Evenness is the measure of statistical validity in the coverage of the genome. A high coverage is evidently preferred, as it means that more of the genome was sampled and there is a smaller chance of missing portions of the genome, however, it is also important to ensure that the coverage was not biased onto some portions of the genome over others: there should be an equal likelihood of a read being generated anywhere on the genome. Datasets which deviate from this pattern are uneven, and

likely to be biased in complex and unpredictable ways. We should instead seek out datasets with a higher level of evenness.

Under the standard statistical assumptions, if the genome is sampled at a uniform rate everywhere, the coverage should follow a Poisson distribution, $\mathcal{P}(k|\lambda)$. However, it is easy to show that the coverage of any platform exhibits a significantly greater dispersion than a Poisson distribution with the correct mean. This is generally interpreted as being indicative that there is not just one rate, λ, at which the genome is sampled, instead there are multiple values, over which the distribution is marginalised [19].

In Supplementary 0, we use this information to generate the following unevenness metric:

$$\mathcal{U} = \frac{\text{Var(coverage)} - \langle\text{coverage}\rangle}{\text{Var(coverage)}}$$

Where Var(coverage) and $\langle$coverage$\rangle$ are the standard statistical variance and mean of the non-zero coverage[2] distributions respectively. This value is zero when the coverage distribution is a perfect Poisson distribution, and is arbitrarily large for distributions which have many values of λ contributing to them. We should therefore favour platforms which generate smaller values of $\mathcal{U}$.

Metric 4: Capability

Capability is the measure of usefulness of the dataset, the ability for the dataset to improve the outcome of a genetic inquiry than would otherwise be achieved without long-range information. A more capable platform produces data which allows the assembly to be vastly improved, and should therefore be preferred.

We measure the capability by comparing the N50 and N90 metrics of a scaffolding with and without the assistance of long range information. The N50

[2] We focus on the non-zero coverge distribution since the designs of the Hi-C and 10x protocols mean higher zero-coverage is to be expected, but the non-zero coverage should be unaffected.

metric is the standard measure of ‘completeness’, it is the length of the shortest continuous sequence such that all longer sequences make up more than 50% of the genome. N90 is defined similarly, but encapsulating 90%. Larger values for Nx are preferred, as this indicates that more of the genome has been grouped into larger, contiguous fragments.

Pair-Linked Reads

Of the three companies that have commercialized Hi-C; Cantata (formerly Dovetail), Arima, and Phase, the most widely applied technologies are OmniC (Cantata) and Arima. In this study, we only carry out analysis on Arima Hi-C reads and comparisons are performed between the two generations of Arima technology (V1 and V2), to reveal characteristics and library improvement by the platform. In total, we obtained three human datasets, two from V2 (NA24385-AJ and NA12878-CEU) and one from V1 (NA12878-CEU; see Table 1).

Hi-C maps are shown in Figure 2 for three human samples by mapping the reads to the human reference assembly GRCh38. In these plots, regions of high density indicate real-space colocation of the genome – though there are some notable deviations from this; in particular, highly repetitive regions can cause spurious over-and-under densities, characterised by a cross-shape running through, i.e. the centromere of each chromosome. To explore the quality of these datasets in more detail, we present Figure 3 which consists of three separate plots showing link-separation distance (association), translocation rates (accessibility) and base coverage (evenness) respectively. Tabulated information regarding the metrics is also presented for accessibility (Table 1) and evenness (Table 3).

Fig. 3A shows how the long-range information is distributed in the Hi-C datasets: as expected we see a peak of very strongly associated regions in the region of 100-500bp (probably due to topologically-associated domains, TADs), and a long power law tail for the three human datasets. For demonstration purposes, we also include an additional dataset – derived from Oak – which demonstrates a strong deviation from the power law structure. In assessing the association demonstrated here, we would say that the oak should be penalised due to this deviation whilst the

human datasets are comparatively much nicer. Aside from this, the plots demonstrate that V2 datasets have more information stored in longer length reads than the V1, and hence have a stronger association.

The Usability metric is shown graphically in Fig. 3b and in more detail in Table 1. We find that the V1 dataset shows a consistently poorer mapping rate, PCR duplication rate and translocation rate over the V2 datasets, resulting in a usability of 0.328, compared to 0.53 and 0.60 for the V2 data, though we do note from Fig. 3B that the difference between human datasets was, on some chromosomes, more pronounced than the difference between platforms.

Fig. 3C and Table 3 show the evenness statistics. Fig. 3B shows the raw coverage data for the Hi-C data, along with a standard Illumina sequencing of the NA12878-CEU sample for comparison: given that the Illumina data has been sequenced more directly, with fewer intervening biochemical alterations, we should expect it to be the "purest" sample. Visually, we can see that this is the case: the Illumina is tightly peaked and resembles a Poisson distribution. The V2 datasets – though sampling to slightly different depths – show a similar "fattened Poisson" distribution, and the V1 data seems to be the least pure sampling, showing a strong overdensity at low base coverage. These observations are carried through by the statistical metric developed in 0: the Illumina data was given a score of 2.7, whilst the V2 platforms both scored approximately 5, and the V1 platform scored 10, indicating a strongly uneven coverage. This would indicate that whilst there is some statistical bias in the V2 data, it is significantly less than that of the V1.

From the information presented here, we would conclude that the V2 platform produces data which robustly outperforms V1, with the two V2 datasets very close together in quality: V2 NA24385-AJ has a slightly higher mean base coverage, but V2 NA12878-CEU scores slightly better on the global accessibility and evenness metric. In the next section, we will demonstrate how Hi-C data can be used to aid Genome Scaffolding, and hence asses the usability of these datasets.

Chain-Linked Reads

Due to their similarity in mechanism and data output, we discuss the 10x and Haplotagging qualities together, presenting an analysis on five datasets, two from from 10x (human and hummingbird) and 3 from Haplotagging (human, rat and oak).

The human 10x dataset was downloaded from the 10x Genomics website and hummingbird dataset is part of the VGP project (see Data availability). The Haplotagging datasets of human, rat and oak were sequenced by the Sanger Institute as part of the Darwin Tree of Life project. We note that, since the data arises from wildly different species, we must take care with our inferences that the differences arise from the choice of platform rather than the choice of species: any strong comparisons should be based primarily on the human samples.

In Figure 4 we see the distribution of molecule lengths for the CLR platforms, which we use as a measure of the association, and in **Error! Reference source not found.** computes the associated barcode collision frequencies for these molecule distributions. We find that all of the platforms produce molecules which have collision rates below 0.1%, and have mean molecule lengths in the region of 50kb. We note that the 10x platforms have a more prominent tail at the high-length end of the distribution, most evident through the N50 values: the 10x N50 values exceed the mean length by 40kb, whilst the haplotagging N50 exceed the mean by only 20kb, indicating the 10x has a stronger tail of high-association data included, even if the bulk of the data has similar associations.

Table 2 shows the PCR duplication rates and the unmapped rates (and hence the accessibility) as well as N50 reads per barcode. We can see clearly that 10x has lower PCR duplications than haplotagging, although the inverse is true for the unmapped rate: this could be largely due to the differing tools used to analyze the datasets (EMA for haplotagging versus LongRanger for 10x), that the 10x data is several years old, whilst the haplotagging is state-of-the-art, and we note that the total accessibility is broadly the same. Probably of more importance is that 10x data exhibits a higher number of N50 reads per barcode than haplotagging.

In Figure 5, we demonstrate the coverage profiles of the CLR datasets, in their raw form in 5a, and, to remove the effects of differing sequencing depths, a normalised form in 5b, where the coverage is given as a fraction of the maximal value. In these figures we can see a number of clear features: firstly, it is clear that both the oak and the rat show extremely strong deviations from an even profile: we hypothesise that this may be due to the effects of highly repetitive regions (that rats are known to possess [20]), which causes some regions of the genome to be erroneously ‘covered’ thousands of times, whilst other regions are deprived of coverage. The human profiles are similar in shape to the Illumina curve. However .

we note that the high coverage end (as with the rat, likely a spurious tail due to over-coverage of repetitive regions) is suppressed relative to the Illumina sample, an indication of the long-range information allowing correct alignment of some repetitive regions. Nevertheless, we do see a stronger bias towards the low coverage end in both the 10x and haplotagging than in Illumina. In contrast, the hummingbird displays a remarkably Poisson-like shape, in large part due to the almost total absence of a repetitive high-coverage tail – likely a feature of a small, non-repetitive genome [21].

These visual conclusions are supported by the unevenness metric in Table 3, where we see both the rat and the oak scoring very poorly (11.5 and 43.6 respectively), the humans scoring between 2 and 9 (illumina: 2.7, haplotagging: 5.1 and 10x: 8.9), and the hummingbird with the lowest score of 1.4. We note that the haplotagging's improved score over the 10x platform is likely a feature of the more powerful suppression of the over-coverage of repetitive regions, rather than of an improvement at the low-coverage end: this might indicate that the haplotagging is more successful in allowing alignment of repetitive regions than 10x.

From the metrics presented here, we conclude that, for base polishing, 10x data is superior to that of Haplotagging due to its slightly higher association strength. However, Haplotagging has a larger number of unique barcodes, resulting in a much lower collision rate, and this means that it is more efficient to handle large number of samples when low coverage data is targeted. In addition, we recall that the statistical properties of the haplotagging platform indicate that haplotagging allows better alignment of highly repetitive regions than 10x.

TELL-Seq

Like haplotagging, TELL-Seq is a promising successor to 10x technology, and so we wish we could have a similar analysis of the platform against the metrics we have formulated here. However, the authors were unable to produce a TELL-Seq library of sufficient quality to provide a viable comparison. We must therefore rely on the literature (i.e. [12]) when discussing the properties of TELL-Seq.

Applications

In this section we briefly outline some of the main applications for long-range data, and discuss how this has been applied in the literature.

Genome scaffolding

Genome scaffolding is the process by which a number of continuous sequences ('contigs') generated from overlapping reads are linked together into a single structure (a scaffold) of known sequences, separated by gaps of unknown sequences but where the length of the gap is relatively well constrained. This forms a critical step in genome assembly [22], but conventional means are both laborious and computationally intensive, though recent advances in long-range sequencing technologies have improved the continuity of genome scaffolds [23], for example, the assembly quality thresholds proposed by Vertebrate Genome Project (VGP) are that contig N50 > 10Mb and the scaffold N50 is the chromosome length [6], indicating that chromosome-scale scaffolding is now routinely possible.

Pair-Linked Reads

The Hi-C protocol provides a fast and lower-cost way of constructing scaffolding from the contigs, given that the spatial information within Hi-C Pair-Linked Reads can identify whether contigs come from the same chromosome and infer the correct orders of the contigs within each chromosome based on the relative proximity between bases in each contig [14]. This technology is widely used to assemble the contigs of eukaryotic genomes into chromosome-scale scaffolds [6, 22], and has recently been applied to assemble the giant and complex genome of Chinese Pine into a chromosome-level assembly [24]. To further improve the quality of genomic assembly, some studies evaluated the different sample preparation kits/protocols and computational programs and identified the optimal conditions for Hi-C scaffolding [25].

To demonstrate how Hi-C data can improve the quality of scaffolding, we applied the above methods to the human genome. 54X HiFi reads from HG002 were downloaded from GIBA and an assembly was obtained using Hifiasm [26] with

contig N50 at 45.1 Mb – this represents the baseline shown in Figure 6A, which is a standard Hi-C proximity map. However, without scaffolding, the proximity map shows a high degree of fragmentation. After removing haplotype duplications, the contigs were further assembled to chromosome-level scaffolds using ~30X Hi-C reads and the YaHS scaffolding tool [27]. The resulting maps for Arima V1 (Figure 6B) and Arima V2 (Fig. 6C) are shown: after scaffolding, chromosome blocks are clearly seen and the fragmentation visibly reduced. Table 4 details how the lengths of the assembled fragments vary after applying the Hi-C data: we find that although V1 produces slightly higher N50 scaffolds and a larger maximum length scaffold, the V2 platform has a higher mean length, indicating a significant reduction in the number of poorly-scaffolded contigs, consistent with our earlier analysis of these platforms.

Detailed instructions on genome assemblies are provided in Supplementary. Assembly pipelines/instructions/recommendations can also be found in VGP, see Rhie, et al., [6].

Chain-Linked Reads

The Chain-Linked strategy provides an effective and less expensive alternative technology than NGS mate pair data for genome scaffolding [22], since the nature of these platforms groups reads by their proximity in the genome (absent any barcode collisions). The information retained from long stretch sequences can be used to link the different contigs to chromosome-level scaffolds, a strategy which is already widely used for genome scaffolding in a variety of complex and polyploid species. For example, in an early study, 10x reads were applied to assist in scaffolding of genome sequence of *Triticum urartu*, the progenitor of the A subgenome of tetraploid and hexaploid wheat [28], and Lee et al. [29] reported that 10x linked reads were successfully used to correct and scaffold the assembly for an allopolyploid rapeseed.

Currently, 10x-based approaches can no longer be used due to the withdrawal of the product, but alternative linked-read technologies have developed based on similar methods, such as Haplotagging and TELL-Seq [12, 13]. As we have seen, the association strength between these platforms is similar, and hence these platforms all offer the ability to produce high quality scaffolding, an assumption which was validated when Chen et al. [12] completed a comprehensive assessment of TELL-Seq using the sequenced data from sample NA12878 and NA24385. The TELL-Seq data

of NA12878 produced a *de novo* assembly with a scaffold N50 of 31.5 Mb, the longest contig 109.2Mb and the longest alignment of 23.6 Mb.

De novo genome assembly

De novo genome assembly is the fundamental process in reconstructing a genome from sequencing reads without a reference sequence [30]. A whole-genome assembly with high level of completeness, continuity and accuracy is the key, which can significantly enhance the reliability of the downstream analyses. In general, the primary step for *de novo* assembly from the collection reads consists of three phases, contig assembly, scaffolding and gap filling. As we saw in 0, both Pair- and Chain-Linked Reads can connect and order contigs into a 'scaffold' in the second phase, however, Chain-Linked Reads also offer additional support for the other procedures involved in *de novo* assembly, due to their inherently more structured nature.

10x technology has been used extensively for the *de novo* assembly of the eukaryotic and prokaryotic genomes [3]. For instance, the first complete genome sequence for the mound-building mouse, *Mus spicilegus*, was generated with 10x reads and resulted in the *de novo* assembly of a 2.50 Gbp genome with a scaffold N50 of 2.27 Mbp [31]. Using 10x data and the Supernova assembler, Ozerov et al. [32] assembled a ~0.8Gb draft genome of the *Silurus glanis*, an important species for freshwater ecosystem balance. It has also been demonstrated that 10x data can be used to assemble high accuracy contigs and scaffolds, even for large, highly similar repetitive sequences, polypoid plant genomes [3, 33].

As comparatively newer technologies, the non-10x CLR platforms have seen less ubiquituous use in *de novo* assembly, though Chen et al. [12] did introduce the TELL-Seq platform by immediately providing *de novo* assembly of bacteria (*Escherichia coli*, *Campylobacter jejuni*, *Rhodobacter sphaeroides*) and humans (NA12878). The human assembly showed "longer aligned contig length and at least 28% and 71% fewer misassemblies than other linked-read or nanopore methods, respectively" [12]. In addition, some attempts have been made to use haplotagging for *de novo* assembly, though the success has been more limited [34].

Although *de novo* genome assembly can be performed by using CLR technologies alone, most current studies adopt a hybrid strategy of multiple

technologies to complete genome assembly. For example, Batra et al. [35] performed a *de novo* genome assembly of the olive baboon using a hybrid sequencing approach of 10x sequencing, Oxford Nanopore sequencing, Illumina paired-end sequencing and Hi-C, which have complementary advantages. Lind et al. [36] generated a high-resolution *de novo* chromosome-scale genome assembly for the Komodo dragon Varanus komodoensis using data from different platforms, including 10x Genomics linked-reads, Oxford Nanopore long reads, PacBio long reads and Bionano optical mapping.

Variation detections

One of the most fundamental goals in genetics is to link genomic variations and the evolution of traits between populations or species. DNA polymorphisms are widespread genomic variations among individuals and include single-nucleotide variants (SNVs), small insertions and deletions (Indels; <50 bp), and structural variations (SVs). Many methods have been proposed to test DNA changes across the genome from different sequencing technologies, but there are still considerable limitations on what can be achieved in SV detection due to technical difficulties of the standard short-read platform. The long-range information provided by CLR and PLR platforms can improve detection for haplotype-specific deletion and large SV [37, 38].

Pair-Linked Reads

Since Hi-C technology detects regions of high interaction probability in a genome, this intrinsically makes it particularly useful for detecting SVs. One of the main advantages of Hi-C is that it can accurately detect SVs with low-depth sequencing data. This feature provides a higher chance of identifying SVs at repetitive regions in complex genomes.

As a result, Hi-C has been demonstrated to be a promising technology to precisely detect SVs, including chromosomal rearrangements and copy number variation in plant and human genomes [39, 40]. In recent years, several research projects have shown the ability of Hi-C to support identifying three-dimensional genome organization alterations as a result of SVs in the human cancer genome [41, 42, 43]. Hi-C has also been applied to screen the complex genomic rearrangements

associated with the development of disease in humans. For example, Melo et al. [44] used Hi-C to investigate the genetic variation that causes developmental disorders, and Hi-C was used to detect multi-megabase polymorphic inversions in wheat and barley [45, 46].

Chain-Linked Reads

Recent work has used SNVs detected by 10x sequencing technology to draw the landscape of meiotic recombination in plant population at the genome-scale resolution [47, 48], and Rommel Fuentes et al. [48] pinpointed meiotic crossovers of interspecific hybrid F1 tomato pollen at the SNV resolution level by using 10x data. This technique also has been a powerful tool for detecting genomic variants associated with human diseases. A number of novel and important SVs associated with metastatic castration-resistant prostate cancer were identified by 10x whole-genome sequencing [49], and CLR sequencing validated the inverted rearrangement in the triple-negative breast cancer sample TN-19 [50]. A 2020 study confirmed that 10x sequencing provides a cost-efficiency way of mining genomic variants at moderate depth and population scale [51], and it was also reported that 10x technology could be used to screen nucleotide resolution of the structural variants linked with potential risk loci in small and rare disease cohorts [52].

Haplotagging is particularly suitable for constructing the original haplotype, and as a result has been successfully applied to construct the genome haplotypes in the two butterfly species, and detect the genetic markers controlling the distinct wing color patterns [13], indicating that haplotagging might be a promising method to identify the superior haplotype alleles in the diverse plant or animal populations for model and non-model species. Bhat et al. [18] thought that this technique would provide important support for haplotype-based breeding for crop improvement.

The utility of the TELL-seq protocol, for detecting genome variations has not been nearly so widely used in the plants or animals, though the study by Chen et al. [12] demonstrated that linked-read data generated by TELL-seq could be used to screen genetic variation using an analysis pipeline developed for the 10x technology.. Althought this means TELL-seq also could be used to detect SVs, the initial study found that it missed some deletions in the NA12878 sample. The authors thought that two factors (the short library insert length and different barcoding chemistry) might be

responsible, and they encourage the research community to further develop and optimize analytical tools to improve the ability to detect SV using linked-read data [12]. More extensive validation studies are therefore needed to prove whether TELL-seq can accurately detect genome-wide variation as an alternate method for the 10x platform.

Other Applications

Phasing

Phasing – the assignment of alleles to either the maternal or paternal haplotype – is another potential application for long-range reads, since even long reads can struggle to accurately identify heterozygosity and correctly assign differences to haplotypes. Along with *de novo* assembly, Chen *et al.* demonstrated how TELL-seq can be used as a powerful tool for phasing the genome :TELL-Seq phasing results on NA12878 and NA24385 samples showed that the highest heterozygous rate is 99.9% and 99.8%, the phasing block N50 is 16.1Mb and 13.4Mb, the longest phasing block is 67.5Mb and 59.2Mb, and adjusted N50 (1.24 Mb), and the lowest switch error rate is 0.04% and 0.08, respectively [12].

Most recently, a study compared the performance and accuracy of genome phasing between Hi-C and 10x Genomics Linked Read in Hanwoo Cattle [53]. The results of this study showed that the phasing strategy with 10x linked-read technology and Long Ranger software displayed the best phasing performance. The best strategy had the highest phasing rate (89.6%), longest adjusted N50 (1.24 Mb), and lowest switch error rate (0.07%). Moreover, the phasing accuracy and yield of the best strategy stayed over 90% for distances up to 4 Mb and 550 Kb, respectively.

Metagenomics

Another application of Chain-Linked Read sequencing technology is assembling high-quality metagenome of microbial species, which is able to improve continuity and accuracy in de novo assembly using barcode information, as comprehensively evaluated by Zhang et al [54]. This study showed that 10x reads significantly improved the metagenome assemblies when compared with Illumina short-reads,

655 although both were outperformed by PacBio CCS long-reads. Due to the low cost and
656 the high base quality, sequencing the metagenomes using Chain-Linked Read
657 technology remains persuasive. Recently, Roodgar et al., [55] explored the
658 longitudinal trajectories of gut microbiome for a single individual using linked-read
659 metagenomic sequencing in 10x Genomics Chromium platform.
660

661 ## Selection of data platforms

662 With rapid development of long read technologies for longer read length and better
663 base accuracy, high profile projects have been lunched such as Vertebrate Genomes
664 Project (VGP) which aims to sequence all the vertebrate species [6] and Darwin Tree
665 of Life (DTOL) which plans to generate de novo assemblies for the 70,000 eukaryotic
666 genomes in Britain and Ireland (https://www.darwintreeoflife.org). More
667 ambitiously the Earth BioGenome Project was proposed to decode ~1.5 million
668 eukaryotic species, including animals, plants and microbiomes [7]. If targeting
669 chromosome-level assemblies, Hi-C data sequencing should be planned, either with
670 Arima V2 or OminC. When sequencing Hi-C in large volumes of data with various
671 speics, the assessment metrics and methods presented in the study could be used for
672 data QC. In the cases where there are choices of platforms, data assessement and
673 comparsions are ensential in order to ensure proper Hi-C libraries are prepared. For
674 small research groups, contracting Hi-C sequencing is one of the options while most
675 sequencing companies provide Hi-C QC reports.
676
677 Chain-Linked reads, such as 10x, Haplotagging and TELL-seq can be used for
678 consensus polishing to improve the quality of genome assembly and enhance the
679 detection of genomic structure variants. We do note that, as of June 30, 2020, 10x
680 Genomics discontinued the sale of Chromium Genome and Exome product lines – the
681 most prominent CLR platform, on which a significant portion of the literature was
682 focused. Various alternatives have been suggested: in this work we studied Sage
683 Science's TELL-Seq platform and Haplotagging. Whilst we found that haplotagging
684 data was in some cases of a higher quality than 10x, haplotagging beads are not (yet)

commercially available, being obtainable from the Chan Lab at the Max Planck lab in Tuebingen only via academic collaboration. Commercial supply of these reagents could make haplotagging a powerful tool, as the beads are potentially inexpensive, which would allow haplotagging to be used widely in genetic population sequencing studies. Additionally, more work on analyzing and processing non-10x data would futher enable the community to make use of these potentially powerful platforms.

Software tools

To date, a large number of tools have been developed to analyze data generated from long-range sequencing technologies [54, 56, 57]. Here, we highlight recent developments in software tools used for genome scaffolding, de novo assembly and variation detection based on the long-range linking information.

Hi-C Analysis Tools

Genome Scaffolding with Hi-C

Several scaffolding methods have been developed for assembling contigs to scaffolds based on Hi-C data, examples of which are shown in Table 5. There are many different approaches which can be taken in designing these tools, which we broadly split into three categories: *deterministic, probabilistic,* and *improver*.

Deterministic tools use algorithms which always return the single result which optimises some underlying metric. Some examples of deterministic algorithms include:

- Heirarchical clustering
  Typically using an agglomerative approach, such as in the early tools LACHESIS [58] and dnaTri [59] (both no longer actively developed), and more recently by ALLHIC [60], a framework particularly designed for scaffolding autopolyploid or heterozygous diploid genomes.
- Best-Neighbour
  Though deterministic, best-neighbour methods return only an approximation to the desired answer, at the benefit of vastly increased speed. 3D-DNA [61] used this approach after correcting the input contigs. The best-neighbour

approach then assembles the contigs into one megascaffold, before it is then cuts to a number of chromosomes on the basis of Hi-C contact matrix.

- Maximal-Matching
  This approach is used in the SALSA1 [62] tool, which first corrects misassemblies derived from the input contig using a low Hi-C mapping rate as the signal for error and then orients and orders the corrected contigs to generate scaffolds using a maximal matching algorithm.
- Novel approaches
  Some more novel solutions include SALSA2 [56], an overhaul of the SALSA1 program that can take advantage of all the interaction information from the Hi-C map to reduce assembly errors using a novel iterative scaffolding method, as well as the newly developed YaHS [27] , which introduced a novel algorithm to establishing the contact matrix to obtain the more accurate inferences of contig joins.

Probabilistic approaches, in contrast, return results which are not exact, but are good approximations to a desired solution where direct computation would be prohibitive. We identify two main classes of probabilistic algorithm.

- Markov Chain Monte Carlo
  MCMC methods are a class of algorithms which attempt to efficiently approximate drawing values from an underlying (unknown) distribution function. This is used by GRAAL [63] which uses a MCMC algorithm to generate scaffolds from the Hi-C data. Recently, Baudry et al. [64] developed instaGRAAL, an upgrade of the GRAAL version, which can be used to assemble large genomes.
- Maximum Likelihood
  Maximum likelihood methods use Bayesian formulations to derive a probability of observing given results, given a hypothesized original state. By optimizing this function, the original state can be inferred. This approach is used by HiRise, the tool developed by Dovetail Genomics for their Hi-C service [65].

Finally, we note a class of tools which we dub *improvers*, these tools do not perform the assembly themselves, but act to improve the quality of assemblies performed using other tools. Examples include HIC-Hiker [66], a probabilistic and dynamic

programming approach which can improve the quality of scaffolds produced by other Hi-C scaffolding software, and the recently developed EndHic [67], which can reduce the error rate of assembly using only the the Hi-C contacts from the end regions of the contigs.

Several studies have evaluated the performance of different scaffolders for scaffolding accuracy [60, 67, 68, 27]. For example, a recent study evaluated the performance of five Hi-C scaffolders including LACHESIS, HiRise, 3D-DNA, SALSA2, and ALLHiC; the results found that the HiRise and LACHESIS display the best performance on average under all tested scripts [68]. However, with all the available software, it remains challenging to correctly assemble large contigs into chromosomes, and manual checking and curation are often necessary. The selection of suitable tools therefore often remains an exercise in trial-and-error by the researcher.

Variation Detection with Hi-C

There exist several computational tools which have been developed to identify SVs from chromatin interaction data. We divide these by the kinds of SV which they can identify.

Tools which can identify Copy Number Variations (CNVs) include HiCNV [41], OneD [69] and HiNT-CNV [70]. Generally speaking, these tools use Bayesian information criteria (and in the case of HiCNV and OneD, Hidden Markov Models, HMM) to identify the location of CNVs. Similar methods can be used to identify interchromosomal translocations – the tools HiCTrans and HiNT-TL are packaged alongside HiCNV and HiNT-CNV respectively [41, 70]. Although the above algorithms were used to screen the SVs within Hi-C data, most of these methods can only detect interchromosomal translocations and long-range intrachromosomal SVs at a low resolution.

Some more specific tools include the HiTea [71] software, developed specifically for identifying mobile transposable element insertions in Hi-C data, as well as NeoLoopFinder [72]; developed for predicting SV-induced chromatin loops, though also capable of detecting complex SVs with Hi-C data. Wang and colleagues [40] have also presented a computational framework, EagleC, which integrates deep-

learning and ensemble-learning strategies to detect a full range of SVs at high resolution.

Overall, there are still strong demands for analysis tools that can use Hi-C data for high-resolution SV detections.

Chain-Linked Read Analysis Tools

Genome Scaffolding with CLR

Generally speaking, most CLR tools should be equally effective, regardless of which CLR platform was used – however due to its prominence, many tools were designed specifically for 10x, and so their applicability to another linked-read platform, including TELL-seq and haplotagging, still need to be further verified. Unlike the Hi-C tools where a wide variety of differing algorithms were used for scaffolding, CLR algorithms broadly follow the same approach: first attempting to unambiguously identify the HMW-DNA fragments each read originated from, before using these fragments as the basis for a scaffolding. The tool fragScaff was first developed for scaffolding the data from contiguity preserving transposase sequencing, but was one of the first tools to receive explicit support for 10x reads [73]. fragScaff uses an explicit threshold metric to determine barcode uniqueness, before constructing and traversing a scaffold graph. ARCS and ARKSare two closely related tools developed by the same team [74, 75]: ARCS is a stand-alone genome scaffolding developed specifically for 10x linked reads, whilst ARKS uses a kmer mapping strategy to align linked reads and contigs to improve computational efficiency, and is an optional additional mode for ARCS. Hiltunen et al. [76] presented a software package ARBitR, which is explicitly designed to work on multiple platforms beyond 10x. The main distinctive feature of the ARBitR is that it consider the overlaps between the involved contigs when splicing, so as to improve the genome scaffolding accuracy.

Other CLR tools include SLR-superscaffolder [77], which uses an inverted top-down approach, and Architect, which uses co-barcoding and paired-end information to improve the contiguity of genome scaffolding [78].

De novo Assembly with CLR

Although there is much mature software that can be applied to de novo assembly of genomes with short-read sequence data [30, 79, 80], only a few comparatively fewer tools have been developed for generating a *de novo* genome from CLR data.

Supernova [16] was developed specifically for *de novo* assembly of genomes that were deeply sequenced using 10x linked-read sequencing platform, by 10x Genomics. Compared to other methods, Supernova can generate phased diploid assemblies over very long distances. Moreover, despite being a 10x product, Supernova can also be used for the data generated on other CLR platforms, such as TELL-seq [12].

Other assembly tools often use a de Bruijin-type approach, for example, cloudSPAdes [81] (an extensible module of the SPAdes assembler) uses CLR[3] data to expand the de Bruijn graph, and can also be applied to metagenomic or hybrid assembly. The Ariadne [82] module uses a novel algorithm, based on de Bruijin Graphs, to handle the barcode deconvolution problem. In their introduction of the TELL-Seq platform, Chen et al. [12] presented TuringAssembler, another de Bruijn graph-based assembler.

Whilst not strictly related to *de* novo assembly, we also note that Bishara et al. presented an assembler, Athena, that use the tag information from linked-read sequencing to improve metagenome assembly [83].

Variation Detection with CLR

A number of tools developed to detect genetic variations in NGS data can also be used on CLR data without significant modification, such as GATK [84], SNVer [85], VarScan [86] and VarDict [87]. However, since these tools do not exploit the long-range information, genome-scale SV detection remains limited, tools which are aware of the long-range information promise much greater detection power.

Long Ranger [38] is the official program developed by 10x Genomics, which can screen variants and SVs, and combines a number of existing tools, such as BWA and GATK, augmented with long-range specific algorithms. GROC-SVs [37] adopt a similar strategy to Long Ranger for identifying SVs, but it performs local assembly on barcoded reads to test high-resolution complex SVs. Recently, a new structural

[3] These platforms use the terminology "Synthetic Long Reads" (SLR), which we have attempted to move away from, preferring instead to refer to them as Chain-Linked Reads.

variant calling software was presented, called LEVIATHAN [88], which can detect SVs in highly fragmented and heterozygosity genomes using similar methods.

A "split molecule" approach has also proven successful, by identifying molecules which are Chain-Linked together, but aligned to disjoint parts of the genome. VALOR [89] has been developed to discover large genomic inversions from linked-read data by an algorithm based on this "split molecule" signature and read pair signature, and an improved version, VALOR2 [90], can identify not only inversions but also other complex SVs involved in segmental duplications, translocations and deletions. LinkedSV [91] also uses split-molecule methods to simultaneously integrate barcode overlapping and enriched fragment endpoints to identify large SVs.

NAIBR [92] identifies SVs by combining the split-molcule approach with a probabilistic model, and similarly, Xia et al. [93] developed the ZoomX tool using probabilistic models SVs signals would be represented in CLRs, meaning ZoomX can detect novel genomic junctions, and hence identify large rearrangements (>200kb).

Conclusions

Here, we discussed the methodologies and applications of long-range, non-local sequencing technologies, focussing on the Pair-Linked Read technology of Hi-C through the Arima V1 and V2 platforms, and the Chain-Linked Read platforms of 10x, Haplotagging and TELL-Seq. Assessing the published literature, we found that Hi-C has been widely used in genome scaffolding to assemble the genome on a chromosomal level, using a wide variety of different algorithmic approaches. Hi-C technology has also been used for assembly curation as well as evaluation and recent efforts have been seen on structural variation detections. The various Chain-Linked Read platforms have been demonstrated to enhance the value of short reads for genome assembly and, in contrast to the PLR platforms, widely used for improved structural variation detection.

We also introduced metrics with which to assess the quality of the sequencing data produced by these platforms, and briefly demonstrated that these metrics provided a robust insight into the ability of the platforms to provide useful genomic

information to researchers finding, for example, that the Arima V2 platform produces significantly higher quality data than the V1 platform.

From our analysis of the existing literature and from our quality metrics, we have found that long-range protocols, including Hi-C and Chain-Linked Read methods, have already been demonstrated to significantly improve the quality of genome assembly and enhance the detection of genomic structure variants, and as NGS technologies and the associated software pipelines continue to develop further, these technologies will continue to move from strength to strength.

We have emphasised throughout this work the distinction between true long-read platforms and the long-range technologies which employ genome partitioning and barcoding to cluster reads into groups providing with much needed long-range information with only a modest cost increase over standard short-read sequencing. Whilst the development of Long-Read technologies would initially seem to make the short-read based technologies discussed here less attractive to researchers, we have demonstrated robustly that non-local information can help supplement Long-Read endeavours, and avoid some of the drawbacks of these emerging technologies, such that a combined long-read/long-range approach remains a cost-effective strategy for complex genome and pan-genome assembly, population genetics, and high-resolution analysis of complex traits.

References

[1] R. Sethi, J. Becker, J. d. Graaf, M. Löwer, M. Suchan, U. Sahin and D. Weber, "Integrative analysis of structural variations using short-reads and linked-reads yields highly specific and sensitive predictions," *PLoS computational biology*, vol. 16, no. 11, p. e1008397, 2020.

[2] S. Goodwin, J. D. McPherson and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, no. 6, pp. 333--351, 2016.

[3] A. Ott, J. C. Schnable, C.-T. Yeh, L. Wu, C. Liu, H.-C. Hu, C. L. Dalgard, S. Sarkar and P. S. Schnable, "Linked read technology for assembling large complex and polyploid genomes," *BMC genomics*, vol. 19, no. 1, pp. 1--15, 2018.

[4] G. A. Logsdon, M. R. Vollger and E. E. Eichler, "Long-read human genome sequencing and its applications," *Nature Reviews Genetics*, vol. 21, no. 10, pp. 597--614, 2020.

[5] S. L. Amarasinghe, S. Su, X. Dong, L. Zappia, M. E. Ritchie and Q. Gouil, "Opportunities and challenges in long-read sequencing data analysis," *Genome biology*, vol. 21, no. 1, pp. 1--16, 2020.

[6] A. Rhie, S. A. McCarthy, O. Fedrigo, J. Damas, G. Formenti, S. Koren, M. Uliano-Silva, W. Chow, A. Fungtammasan, J. Kim and others, "Towards complete and error-free genome assemblies of all vertebrate species," *Nature*, vol. 592, no. 7856, pp. 737--746, 2021.

[7] H. A. Lewin, G. E. Robinson, W. J. Kress, W. J. Baker, J. Coddington, K. A. Crandall, R. Durbin, S. V. Edwards, F. Forest, M. T. P. Gilbert and others, "Earth BioGenome Project: Sequencing life for the future of life," *Proceedings of the National Academy of Sciences*, vol. 115, no. 17, pp. 4325--4333, 2018.

[8] S. Selvaraj, J. R Dixon, V. Bansal and B. Ren, "Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing," *Nature biotechnology*, vol. 31, no. 12, pp. 1111--1118, 2013.

[9] F. J. Sedlazeck, H. Lee, C. A. Darby and M. C. Schatz, "Piercing the dark matter: bioinformatics of long-range sequencing and mapping," *Nature Reviews Genetics*, vol. 19, no. 6, pp. 329--346, 2018.

[10] A. Bankevich and P. A. Pevzner, "TruSPAdes: barcode assembly of TruSeq synthetic long reads," *Nature methods*, vol. 13, no. 3, pp. 248--250, 2016.

[11] I. Wu, H. S. Kim and T. Ben-Yehezkel, "A single-molecule long-read survey of human transcriptomes using LoopSeq synthetic long read sequencing," *bioRxiv*, p. 532135, 2019.

[12] Z. Chen, L. Pham, T.-C. Wu, G. Mo, Y. Xia, P. L. Chang, D. Porter, T. Phan, H. Che, H. Tran and others, "Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information," *Genome research*, vol. 30, no. 6, pp. 898--909, 2020.

[13] J. I. Meier, P. A. Salazar, M. Ku\vcka, R. W. Davies, A. Dréau, I. Aldás, O. Box Power, N. J. Nadeau, J. R. Bridle, C. Rolian and others, "Haplotype tagging reveals parallel formation of hybrid races in two butterfly species," *Proceedings of the National Academy of Sciences*, vol. 118, no. 25, p. e2015005118, 2021.

[14] E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner and others, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *science*, vol. 326, no. 5950, pp. 289--293, 2009.

[15] J. O. Korbel and C. Lee, "Genome assembly and haplotyping with Hi-C," *Nature biotechnology*, vol. 31, no. 12, pp. 1099--1101, 2013.

[16] N. I. Weisenfeld, V. Kumar, P. Shah, D. M. Church and D. B. Jaffe, "Direct determination of diploid genome sequences," *Genome research*, vol. 27, no. 5, pp. 757--767, 2017.

[17] 1. Genomics, "Linked Reads, Retrieved from https://www.10xgenomics.com/products/linked-reads on 2020-10-17.," 2020.

[18] J. A. Bhat, D. Yu, A. Bohra, S. A. Ganie and R. K. Varshney, "Features and applications of haplotypes in crop breeding," *Communications biology*, vol. 4, no. 1, pp. 1--12, 2021.

[19] M. S. Lindner, M. Kollock, F. Zickmann and B. Y. Renard, "Analyzing genome coverage profiles with applications to quality control in metagenomics," *Bioinformatics*, vol. 29, no. 10, pp. 1260--1267, 2013.

[20] S. K. Shore, L. T. Bacheler, J. Kimball de Riel, L. R. Barrows and M. Lynch, "Cloning and characterization of a rat-specific repetitive DNA sequence," *Gene*, vol. 45, no. 1, pp. 87-93, 1986.

[21] T. R. Gregory, C. B. Andrews, J. A. McGuire and C. C. Witt, "The smallest avian genomes are found in hummingbirds," *Proceedings of the Royal Society B: Biological Sciences*, vol. 276, no. 1674, pp. 3753--3757, 2009.

[22] J. Luo, Y. Wei, M. Lyu, Z. Wu, X. Liu, H. Luo and C. Yan, "A comprehensive review of scaffolding methods in genome assembly," *Briefings in Bioinformatics*, vol. 22, no. 5, p. bbab033, 2021.

[23] E. S. Rice, R. E. Green and others, "New approaches for genome assembly and scaffolding," *Annu Rev Anim Biosci*, vol. 7, no. 1, pp. 17--40, 2019.

[24] S. Niu, J. Li, W. Bo, W. Yang, A. Zuccolo, S. Giacomello, X. Chen, F. Han, J. Yang, Y. Song and others, "The Chinese pine genome and methylome unveil key features of conifer evolution," *Cell*, vol. 185, no. 1, pp. 204--217, 2022.

[25] K. Yamaguchi, M. Kadota, O. Nishimura, Y. Ohishi, Y. Naito and S. Kuraku, "Technical considerations in Hi-C scaffolding and evaluation of chromosome-scale genome assemblies," *Molecular Ecology*, vol. 30, no. 23, pp. 5923--5934, 2021.

[26] H. Cheng, G. T. Concepcion, X. Feng, H. Zhang and H. Li, "Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm," *Nature methods*, vol. 18, no. 2, pp. 170--175, 2021.

[27] C. Zhou, S. A. McCarthy and R. Durbin, "YaHS: yet another Hi-C scaffolding tool," *bioRxiv*, 2022.

[28] H.-Q. Ling, B. Ma, X. Shi, H. Liu, L. Dong, H. Sun, Y. Cao, Q. Gao, S. Zheng, Y. Li and others, "Genome sequence of the progenitor of wheat A subgenome Triticum urartu," *Nature*, vol. 557, no. 7705, pp. 424--428, 2018.

[29] H. Lee, H. S. Chawla, C. Obermeier, F. Dreyer, A. Abbadi and R. Snowdon, "Chromosome-scale assembly of winter oilseed rape Brassica napus," *Frontiers in plant science*, vol. 11, p. 496, 2020.

[30] J.-i. Sohn and J.-W. Nam, "The present and future of de novo whole-genome assembly," *Briefings in bioinformatics*, vol. 19, no. 1, pp. 23--40, 2018.

[31] M. B. Couger, L. Arévalo and P. Campbell, "A high quality genome for Mus spicilegus, a close relative of house mice with unique social and ecological adaptations," *G3: Genes, Genomes, Genetics*, vol. 8, no. 7, pp. 2145--2152, 2018.

[32] M. Y. Ozerov, M. Flaj\vshans, K. Noreikiene, A. Vasemägi and R. Gross, "Draft genome assembly of the freshwater apex predator wels catfish (Silurus glanis) using linked-read sequencing," *G3: Genes, Genomes, Genetics*, vol. 10, no. 11, pp. 3897--3906, 2020.

[33] A. M. Hulse-Kemp, S. Maheshwari, K. Stoffel, T. A. Hill, D. Jaffe, S. R. Williams, N. Weisenfeld, S. Ramakrishnan, V. Kumar, P. Shah and others, "Reference quality assembly of the 3.5-Gb genome of Capsicum annuum from a single linked-read library," *Horticulture research*, vol. 5, 2018.

[34] F. Chan, *Private Communications*, 2022.

[35] S. S. Batra, M. Levy-Sakin, J. Robinson, J. Guillory, S. Durinck, T. P. Vilgalys, P.-Y. Kwok, L. A. Cox, S. Seshagiri, Y. S. Song and others, "Accurate assembly of the olive baboon (Papio anubis) genome using long-read and Hi-C data," *Gigascience*, vol. 9, no. 12, p. giaa134, 2020.

[36] A. L. Lind, Y. Y. Lai, Y. Mostovoy, A. K. Holloway, A. Iannucci, A. C. Mak, M. Fondi, V. Orlandini, W. L. Eckalbar, M. Milan and others, "Genome of the Komodo dragon reveals adaptations in the cardiovascular and chemosensory systems of monitor lizards," *Nature ecology \& evolution*, vol. 3, no. 8, pp. 1241--1252, 2019.

[37] N. Spies, Z. Weng, A. Bishara, J. McDaniel, D. Catoe, J. M. Zook, M. Salit, R. B. West, S. Batzoglou and A. Sidow, "Genome-wide reconstruction of complex structural variants using read clouds," *Nature methods*, vol. 14, no. 9, pp. 915--920, 2017.

[38] P. Marks, S. Garcia, A. M. Barrio, K. Belhocine, J. Bernate, R. Bharadwaj, K. Bjornson, C. Catalanotti, J. Delaney, A. Fehr and others, "Resolving the full spectrum of human genome variation using Linked-Reads," *Genome research*, vol. 29, no. 4, pp. 635--645, 2019.

[39] L. Harewood, K. Kishore, M. D. Eldridge, S. Wingett, D. Pearson, S. Schoenfelder, V. P. Collins and P. Fraser, "Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours," *Genome biology*, vol. 18, no. 1, pp. 1--11, 2017.

[40] X. Wang, Y. Luan and F. Yue, "EagleC: A deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps," *Science advances*, vol. 8, no. 24, p. eabn9215, 2022.

[41] A. Chakraborty and F. Ay, "Identification of copy number variations and translocations in cancer cells from Hi-C data," *Bioinformatics*, vol. 34, no. 2, pp. 338--345, 2018.

[42] J. R. Dixon, J. Xu, V. Dileep, Y. Zhan, F. Song, V. T. Le, G. G. Yard\imc\i, A. Chakraborty, D. V. Bann, Y. Wang and others, "Integrative detection and analysis of structural variation in cancer genomes," *Nature genetics*, vol. 50, no. 10, pp. 1388--1398, 2018.

[43] E. C. Jacobson, R. S. Grand, J. K. Perry, M. H. Vickers, A. L. Olins, D. E. Olins and J. M. O'Sullivan, "Hi-C detects novel structural variants in HL-60 and HL-60/S4 cell lines," *Genomics*, vol. 112, no. 1, pp. 151--162, 2020.

[44] U. S. Melo, R. Schöpflin, R. Acuna-Hidalgo, M. A. Mensah, B. Fischer-Zirnsak, M. Holtgrewe, M.-K. Klever, S. Türkmen, V. Heinrich, I. D. Pluym and others, "Hi-C identifies complex genomic rearrangements and TAD-shuffling in developmental diseases," *The American Journal of Human Genetics*, vol. 106, no. 6, pp. 872--884, 2020.

[45] A. Himmelbach, A. Ruban, I. Walde, H. \vSimková, J. Dole\vzel, A. Hastie, N. Stein and M. Mascher, "Discovery of multi-megabase polymorphic inversions by chromosome conformation capture sequencing in large-genome plant species," *The Plant Journal*, vol. 96, no. 6, pp. 1309--1316, 2018.

[46] S. Walkowiak, L. Gao, C. Monat, G. Haberer, M. T. Kassa, J. Brinton, R. H. Ramirez-Gonzalez, M. C. Kolodziej, E. Delorean, D. Thambugala and others, "Multiple wheat genomes reveal global variation in modern breeding," *Nature*, vol. 588, no. 7837, pp. 277--283, 2020.

[47] H. Sun, B. A. Rowan, P. J. Flood, R. Brandt, J. Fuss, A. M. Hancock, R. W. Michelmore, B. Huettel and K. Schneeberger, "Linked-read sequencing of gametes allows efficient genome-wide analysis of meiotic recombination," *Nature communications*, vol. 10, no. 1, pp. 1--9, 2019.

[48] R. Rommel Fuentes, T. Hesselink, R. Nieuwenhuis, L. Bakker, E. Schijlen, W. van Dooijeweert, S. Diaz Trivino, J. R. de Haan, G. Sanchez Perez, X. Zhang and others, "Meiotic recombination profiling of interspecific hybrid F1 tomato pollen by linked read sequencing," *The Plant Journal*, vol. 102, no. 3, pp. 480--492, 2020.

[49] S. R. Viswanathan, G. Ha, A. M. Hoff, J. A. Wala, J. Carrot-Zhang, C. W. Whelan, N. J. Haradhvala, S. S. Freeman, S. C. Reed, J. Rhoades and others, "Structural alterations driving castration-resistant prostate cancer revealed by linked-read genome sequencing," *Cell*, vol. 174, no. 2, pp. 433--447, 2018.

[50] M. Kawazu, S. Kojima, T. Ueno, Y. Totoki, H. Nakamura, A. Kunita, W. Qu, J. Yoshimura, M. Soda, T. Yasuda and others, "Integrative analysis of genomic alterations in triple-negative breast cancer in association with homologous recombination deficiency," *PLoS genetics*, vol. 13, no. 6, p. e1006853, 2017.

[51] D. Lutgen, R. Ritter, R.-A. Olsen, H. Schielzeth, J. Gruselius, P. Ewels, J. T. García, H. Shirihai, M. Schweizer, A. Suh and others, "Linked-read sequencing enables haplotype-resolved resequencing at population scale," *Molecular ecology resources*, vol. 20, no. 5, pp. 1311--1322, 2020.

[52] K.-T. Tan, H. Kim, J. Carrot-Zhang, Y. Zhang, W. J. Kim, G. Kugener, J. A. Wala, T. P. Howard, Y.-Y. Chi, R. Beroukhim and others, "Haplotype-resolved germline and somatic alterations in renal medullary carcinomas," *Genome medicine*, vol. 13, no. 1, pp. 1--13, 2021.

[53] K. Srikanth, J.-E. Park, D. Lim, J. Cha, S.-R. Cho, I.-C. Cho and W. Park, "A comparison between hi-C and 10X genomics linked read sequencing for whole genome phasing in Hanwoo cattle," *Genes*, vol. 11, no. 3, p. 332, 2020.

[54] L. Zhang, X. Fang, H. Liao, Z. Zhang, X. Zhou, L. Han, Y. Chen, Q. Qiu and S. C. Li, "A comprehensive investigation of metagenome assembly by linked-read sequencing," *Microbiome*, vol. 8, no. 1, pp. 1--11, 2020.

[55] M. Roodgar, B. H. Good, N. R. Garud, S. Martis, M. Avula, W. Zhou, S. M. Lancaster, H. Lee, A. Babveyh, S. Nesamoney and others, "Longitudinal linked-read sequencing reveals ecological and evolutionary responses of a human gut microbiome during antibiotic treatment," *Genome research*, vol. 31, no. 8, pp. 1433--1446, 2021.

[56] J. Ghurye and M. Pop, "Modern technologies and algorithms for scaffolding assembled genomes," *PLoS computational biology*, vol. 15, no. 6, p. e1006994, 2019.

[57] X. Liao, M. Li, Y. Zou, F.-X. Wu, J. Wang and others, "Current challenges and solutions of de novo assembly," *Quantitative Biology*, vol. 7, no. 2, pp. 90--109, 2019.

[58] J. N. Burton, A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman and J. Shendure, "Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions," *Nature biotechnology*, vol. 31, no. 12, pp. 1119--1125, 2013.

[59] N. Kaplan and J. Dekker, "High-throughput genome scaffolding from in vivo DNA interaction frequency," *Nature biotechnology*, vol. 31, no. 12, pp. 1143--1147, 2013.

[60] X. Zhang, S. Zhang, Q. Zhao, R. Ming and H. Tang, "Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data," *Nature plants*, vol. 5, no. 8, pp. 833--845, 2019.

[61] O. Dudchenko, S. S. Batra, A. D. Omer, S. K. Nyquist, M. Hoeger, N. C. Durand, M. S. Shamim, I. Machol, E. S. Lander, A. P. Aiden and others, "De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds," *Science*, vol. 356, no. 6333, pp. 92--95, 2017.

[62] J. Ghurye, M. Pop, S. Koren, D. Bickhart and C.-S. Chin, "Scaffolding of long read assemblies using long range contact information," *BMC genomics*, vol. 18, no. 1, pp. 1--11, 2017.

[63] H. Marie-Nelly, M. Marbouty, A. Cournac, J.-F. Flot, G. Liti, D. P. Parodi, S. Syan, N. Guillén, A. Margeot, C. Zimmer and others, "High-quality genome (re) assembly using chromosomal contact data," *Nature communications*, vol. 5, no. 1, pp. 1--10, 2014.

[64] L. Baudry, N. Guiglielmoni, H. Marie-Nelly, A. Cormier, M. Marbouty, K. Avia, Y. L. Mie, O. Godfroy, L. Sterck, J. M. Cock and others, "instaGRAAL: chromosome-level quality scaffolding of genomes using a proximity ligation-based scaffolder," *Genome biology*, vol. 21, no. 1, pp. 1--22, 2020.

[65] N. H. Putnam, B. L. O'Connell, J. C. Stites, B. J. Rice, M. Blanchette, R. Calef, C. J. Troll, A. Fields, P. D. Hartley, C. W. Sugnet and others, "Chromosome-scale shotgun assembly using an in vitro method for long-range linkage," *Genome research*, vol. 26, no. 3, pp. 342--350, 2016.

[66] R. Nakabayashi and S. Morishita, "HiC-Hiker: a probabilistic model to determine contig orientation in chromosome-length scaffolds with Hi-C," *Bioinformatics*, vol. 36, no. 13, pp. 3966--3974, 2020.

[67] S. Wang, H. Wang, F. Jiang, A. Wang, H. Liu, H. Zhao, B. Yang, D. Xu, Y. Zhang and W. Fan, "EndHiC: assemble large contigs into chromosomal-level scaffolds using the Hi-C links from contig ends," *arXiv preprint arXiv:2111.15411*, 2021.

[68] A. Sur, W. S. Noble and P. J. Myler, "A benchmark of Hi-C scaffolders using reference genomes and de novo assemblies," *bioRxiv*, 2022.

[69] E. Vidal, F. le Dily, J. Quilez, R. Stadhouders, Y. Cuartero, T. Graf, M. A. Marti-Renom, M. Beato and G. J. Filion, "OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes," *Nucleic acids research*, vol. 46, no. 8, pp. e49--e49, 2018.

[70] S. Wang, S. Lee, C. Chu, D. Jain, P. Kerpedjiev, G. M. Nelson, J. M. Walsh, B. H. Alver and P. J. Park, "HiNT: a computational method for detecting copy number variations and translocations from Hi-C data," *Genome biology*, vol. 21, no. 1, pp. 1--15, 2020.

[71] D. Jain, C. Chu, B. H. Alver, S. Lee, E. A. Lee and P. J. Park, "HiTea: a computational pipeline to identify non-reference transposable element insertions in Hi-C data," *Bioinformatics*, vol. 37, no. 8, pp. 1045--1051, 2021.

[72] X. Wang, J. Xu, B. Zhang, Y. Hou, F. Song, H. Lyu and F. Yue, "Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes," *Nature methods*, vol. 18, no. 6, pp. 661--668, 2021.

[73] A. Adey, J. O. Kitzman, J. N. Burton, R. Daza, A. Kumar, L. Christiansen, M. Ronaghi, S. Amini, K. L. Gunderson, F. J. Steemers and others, "In vitro, long-range sequence information for de novo genome assembly via transposase contiguity," *Genome research*, vol. 24, no. 12, pp. 2041--2049, 2014.

[74] L. Coombe, J. Zhang, B. P. Vandervalk, J. Chu, S. D. Jackman, I. Birol and R. L. Warren, "ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers," *BMC bioinformatics*, vol. 19, no. 1, pp. 1--10, 2018.

[75] S. Yeo, L. Coombe, R. L. Warren, J. Chu and I. Birol, "ARCS: scaffolding genome drafts with linked reads," *Bioinformatics*, vol. 34, no. 5, pp. 725--731, 2018.

[76] M. Hiltunen, M. Ryberg and H. Johannesson, "ARBitR: an overlap-aware genome assembly scaffolder for linked reads," *Bioinformatics*, vol. 37, no. 15, pp. 2203--2205, 2021.

[77] L. Guo, M. Xu, W. Wang, S. Gu, X. Zhao, F. Chen, O. Wang, X. Xu, I. Seim, G. Fan and others, "SLR-superscaffolder: a de novo scaffolding tool for synthetic long reads using a top-to-bottom scheme," *BMC bioinformatics*, vol. 22, no. 1, pp. 1--16, 2021.

[78] V. Kuleshov, M. P. Snyder and S. Batzoglou, "Genome assembly from synthetic long read clouds," *Bioinformatics*, vol. 32, no. 12, pp. i216--i224, 2016.

[79] K. Paszkiewicz and D. J. Studholme, "De novo assembly of short sequence reads," *Briefings in bioinformatics*, vol. 11, no. 5, pp. 457--472, 2010.

[80] A. R. Khan, M. T. Pervez, M. E. Babar, N. Naveed and M. Shoaib, "A comprehensive study of de novo genome assemblers: current challenges and future prospective," *Evolutionary Bioinformatics*, vol. 14, p. 1176934318758650, 2018.

[81] I. Tolstoganov, A. Bankevich, Z. Chen and P. A. Pevzner, "cloudSPAdes: assembly of synthetic long reads using de Bruijn graphs," *Bioinformatics*, vol. 35, no. 14, pp. i61--i70, 2019.

[82] L. Mak, D. Meleshko, D. C. Danko, W. N. Barakzai, N. Belchikov and I. Hajirasouliha, "Ariadne: Synthetic Long Read Deconvolution Using Assembly Graphs," *BioRxiv*, pp. 2021--05, 2022.

[83] A. Bishara, E. L. Moss, M. Kolmogorov, A. E. Parada, Z. Weng, A. Sidow, A. E. Dekas, S. Batzoglou and A. S. Bhatt, "High-quality genome sequences of uncultured microbes by assembly of read clouds," *Nature biotechnology*, vol. 36, no. 11, pp. 1067--1075, 2018.

[84] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. Del Angel, M. A. Rivas, M. Hanna and others, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nature genetics*, vol. 43, no. 5, pp. 491--498, 2011.

[85] Z. Wei, W. Wang, P. Hu, G. J. Lyon and H. Hakonarson, "SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data," *Nucleic acids research*, vol. 39, no. 19, pp. e132--e132, 2011.

[86] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding and R. K. Wilson, "VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing," *Genome research*, vol. 22, no. 3, pp. 568--576, 2012.

[87] Z. Lai, A. Markovets, M. Ahdesmaki, B. Chapman, O. Hofmann, R. McEwen, J. Johnson, B. Dougherty, J. C. Barrett and J. R. Dry, "VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research," *Nucleic acids research*, vol. 44, no. 11, pp. e108--e108, 2016.

[88] P. Morisse, F. Legeai and C. Lemaitre, "LEVIATHAN: efficient discovery of large structural variants by leveraging long-range information from Linked-Reads data," *bioRxiv*, 2021.

[89] M. Eslami Rasekh, G. Chiatante, M. Miroballo, J. Tang, M. Ventura, C. T. Amemiya, E. E. Eichler, F. Antonacci and C. Alkan, "Discovery of large genomic inversions using long range information," *BMC genomics*, vol. 18, no. 1, pp. 1--12, 2017.

[90] F. Karao\uglano\uglu, C. Ricketts, E. Ebren, M. E. Rasekh, I. Hajirasouliha and C. Alkan, "VALOR2: characterization of large-scale structural variants using linked-reads," *Genome biology*, vol. 21, no. 1, pp. 1--11, 2020.

[91] L. Fang, C. Kao, M. V. Gonzalez, F. A. Mafra, R. Pellegrino da Silva, M. Li, S.-S. Wenzel, K. Wimmer, H. Hakonarson and K. Wang, "LinkedSV for detection of mosaic structural variants from linked-read exome and genome sequencing data," *Nature communications*, vol. 10, no. 1, pp. 1--15, 2019.

[92] R. Elyanow, H.-T. Wu and B. J. Raphael, "Identifying structural variants using linked-read sequencing data," *Bioinformatics*, vol. 34, no. 2, pp. 353--360, 2018.

[93] L. C. Xia, J. M. Bell, C. Wood-Bouwens, J. J. Chen, N. R. Zhang and H. P. Ji, "Identification of large rearrangements in cancer genomes with barcode linked reads," *Nucleic acids research*, vol. 46, no. 4, pp. e19--e19, 2018.

Acknowledgements

We thank Dr Anthony Schmitt, Arima Genomics for providing data to evaluate Hi-C technology and its developments. We are also grateful to Professor Yingguang Frank Chan, Friedrich Miescher Laboratory of the Max Planck Society, for providing much needed technical support on haplotagging library preparation and sequencing. Funding: L.J. is supported by grant 32070601 from the National Natural Science Foundation of China. Z.N., M.Q., J.F-G., K.O. and E.M. G.are supported by the Wellcome Trust (WT206194). Author contributions: L.J., M.Q. and Z.N. proposed and designed the project. M.Q., K.O. and E.M.G. made the haplotagging sequencing libraries and produced the data. Z.N. performed data analysis, while J.F-G. developed the statistic model for coverage assessment and barcode collisions. H.W., X.S. and L.J. drew Figure 1. L.J., J.F-G. and Z.N. wrote the paper. All the authors read and approved the final manuscript.

Data availability

Hi-C reads were sequenced and provide by Arima Genomics and have been submitted to NCBI under BioProjectID PRJNAxxxxx. The 10x human genome reads were produced by Genome-in-a-bottle and can be downloaded from ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/NA12878/10Xgenomics_ChromiumGenome_LongRanger2.0_06202016/NA12878_GRCh38.bam. The 10x hummingbird dataset has been archived on NCBI/EBI BioProject under

915 accession PRJNA489243. All the haplotagging datasets are part of Darwin Tree of
916 Life project and have been submitted to NCBI with BioProjectID PRJNA435xxx. No
917 sequencing effort has been paid on the TELL-Seq platform and the contents presented
918 in this study are based on the TELL-Seq paper, Chen *et al.* [12].

919

List of Figures

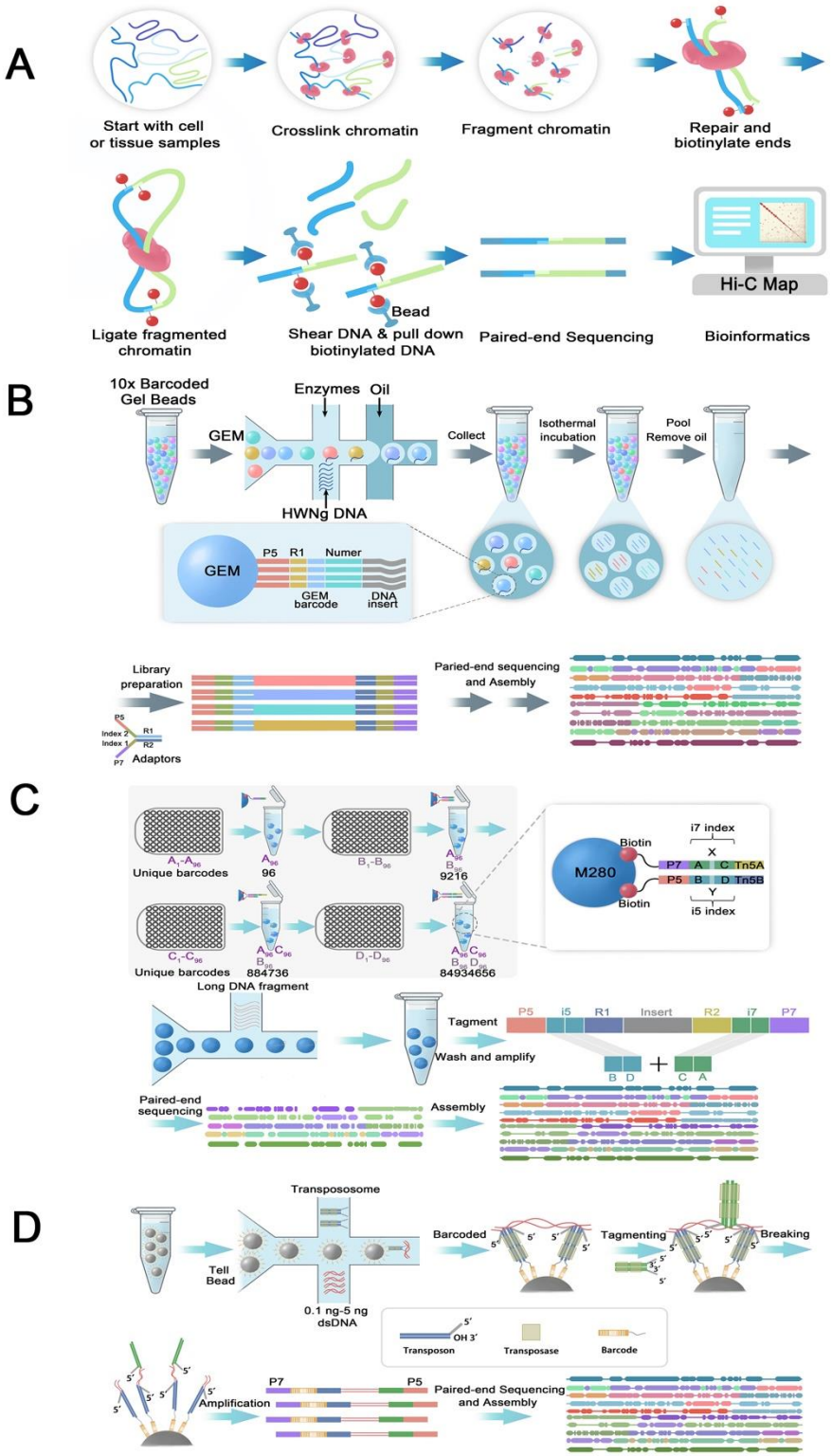
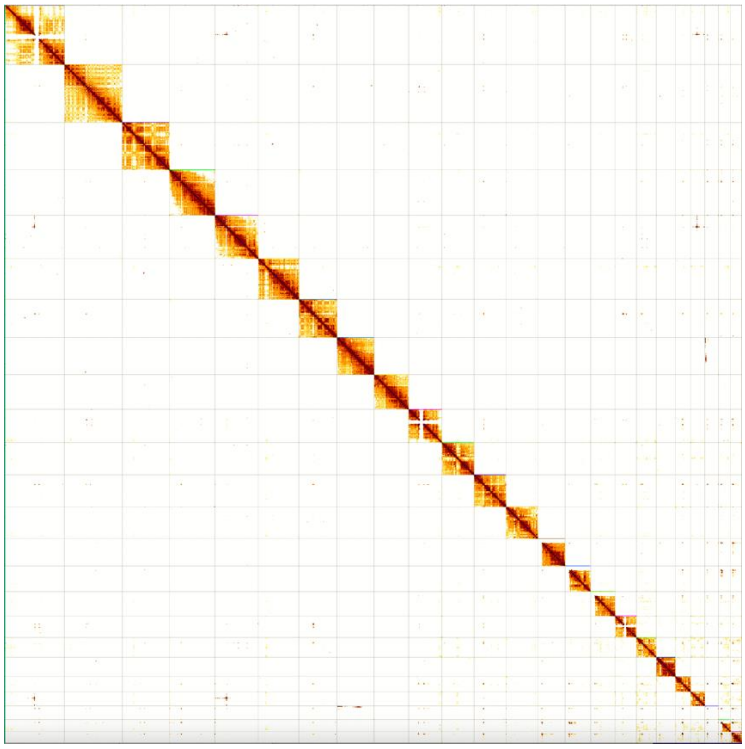
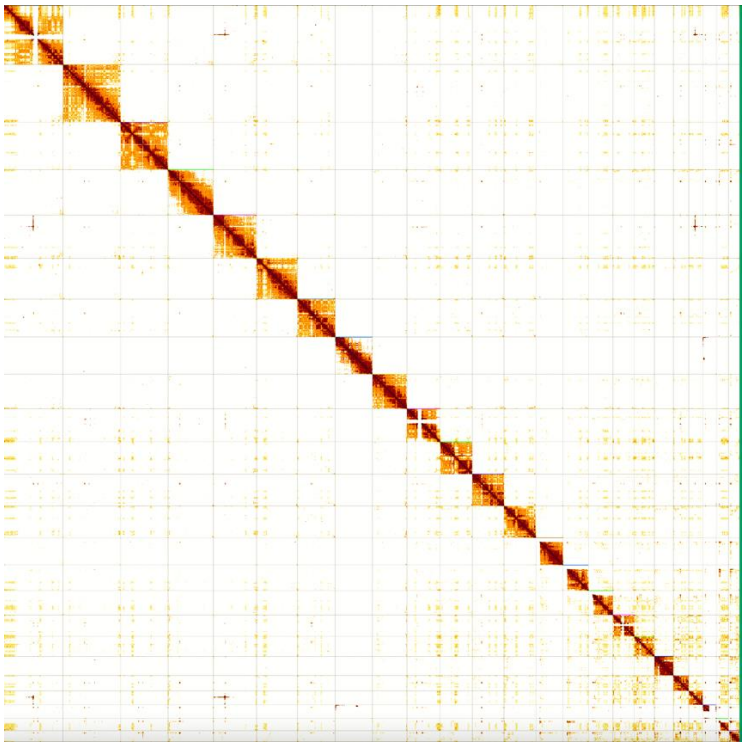

A
Start with cell or tissue samples
Crosslink chromatin
Fragment chromatin
Repair and biotinylate ends
Ligate fragmented chromatin
Shear DNA & pull down biotinylated DNA
Bead
Paired-end Sequencing
Hi-C Map
Bioinformatics
B
10x Barcoded Gel Beads
GEM
Enzymes
Oil
HWNg DNA
Collect
Isothermal incubation
Pool
Remove oil
GEM
P5
R1
Numer
GEM barcode
DNA insert
Library preparation
Adaptors
Paried-end sequencing and Asembly
C
Unique barcodes
96
9216
Unique barcodes
884736
84934656
M280
Biotin
i7 index
i5 index
Tn5A
Tn5B
Long DNA fragment
Tagment
Wash and amplify
Paired-end sequencing
Assembly
P5
i5
R1
Insert
R2
i7
P7
D
Transpososome
Tell Bead
0.1 ng-5 ng dsDNA
Barcoded
Tagmenting
Breaking
Transposon
Transposase
Barcode
Amplification
Paired-end Sequencing and Assembly

Figure 1

949 A B C



950



951

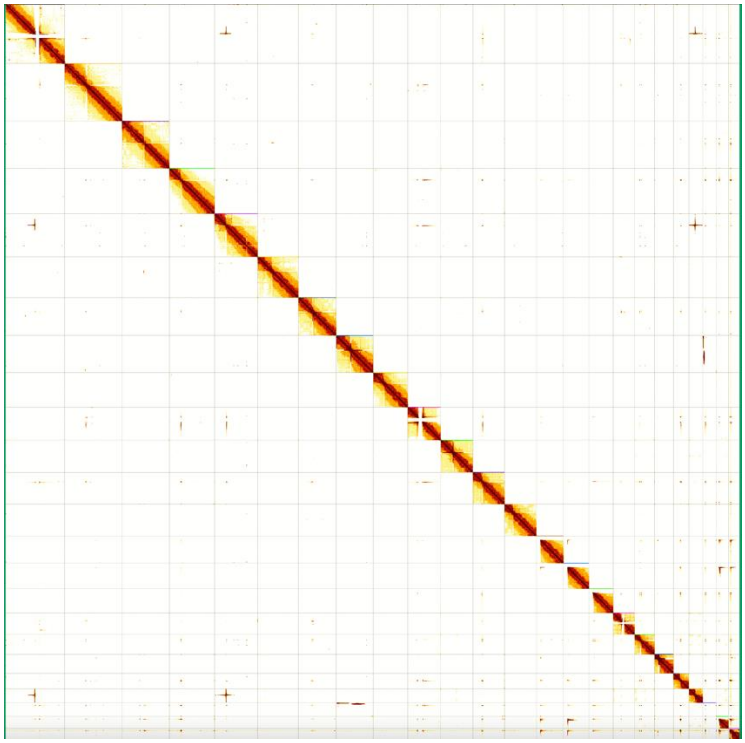


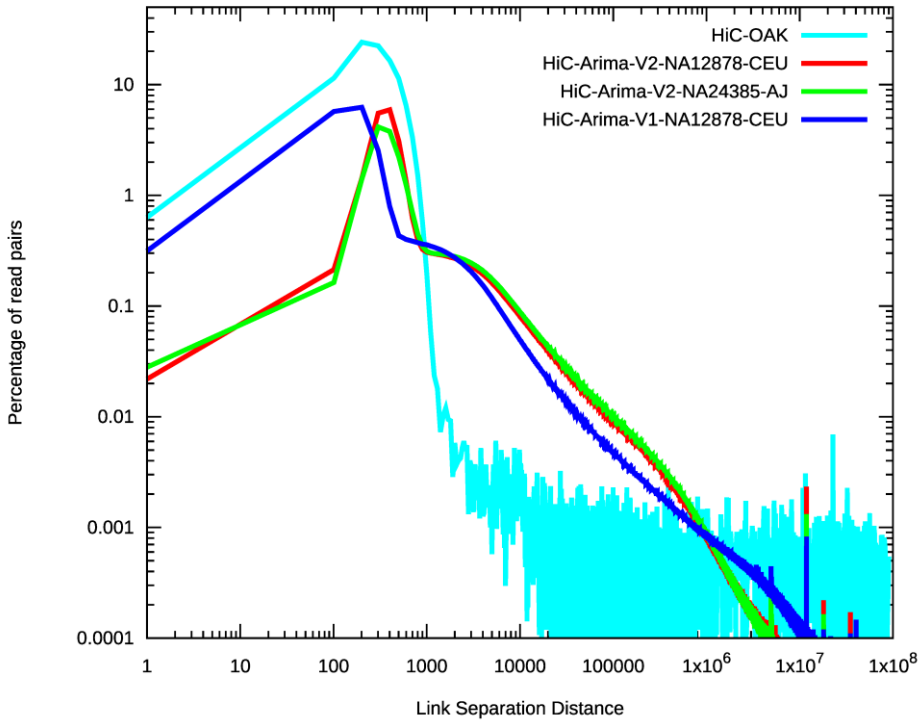
Figure 2

960

961

962 A B C

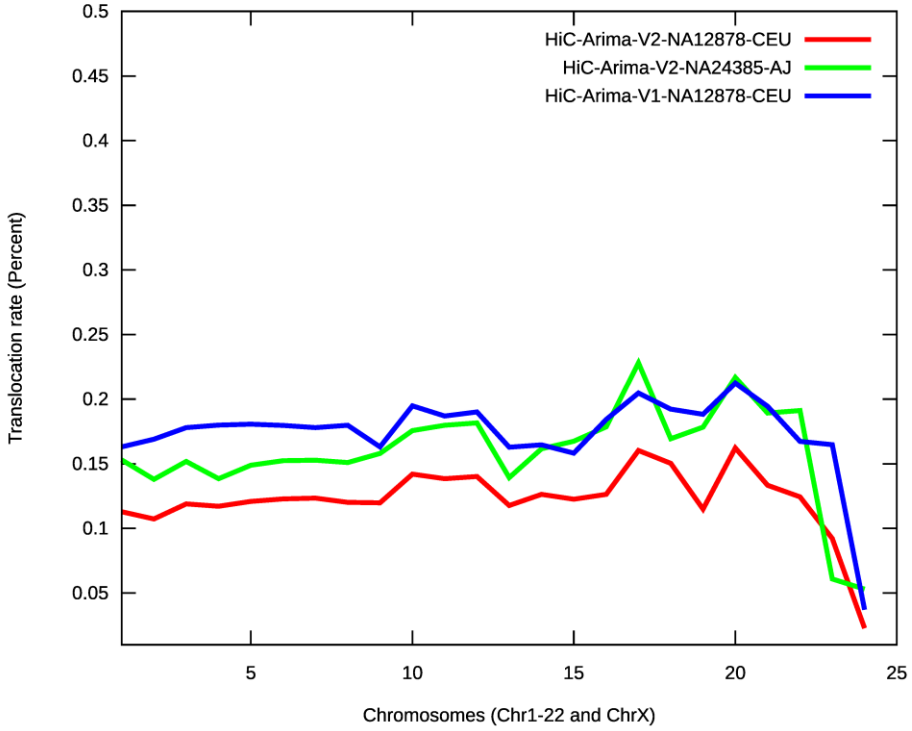
963



964

965

966

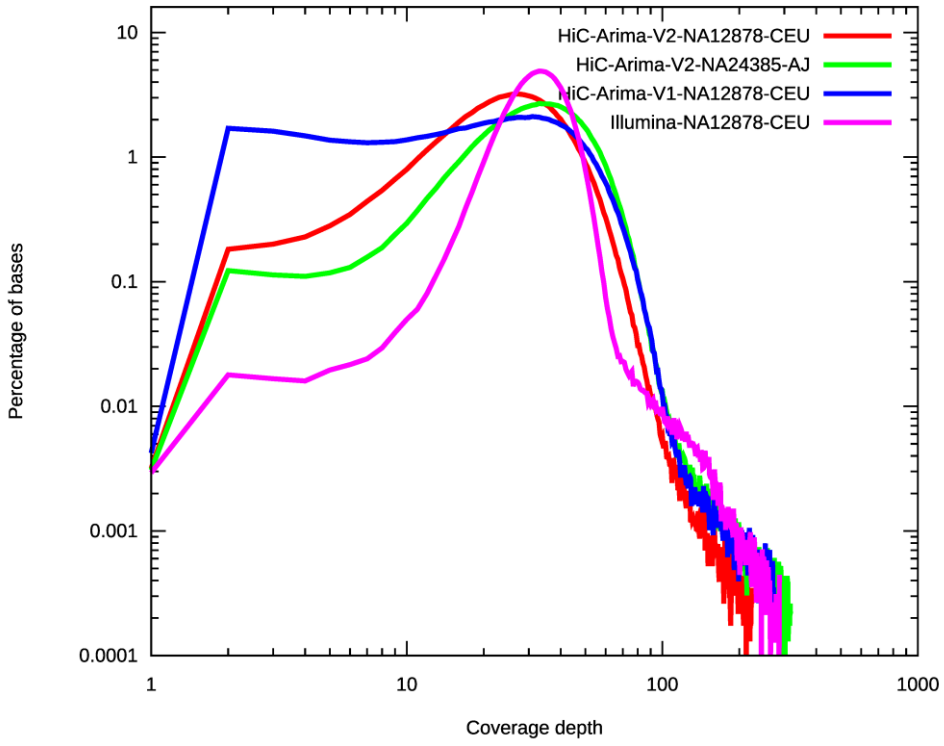


967

968

969

970



971

972 Figure 3

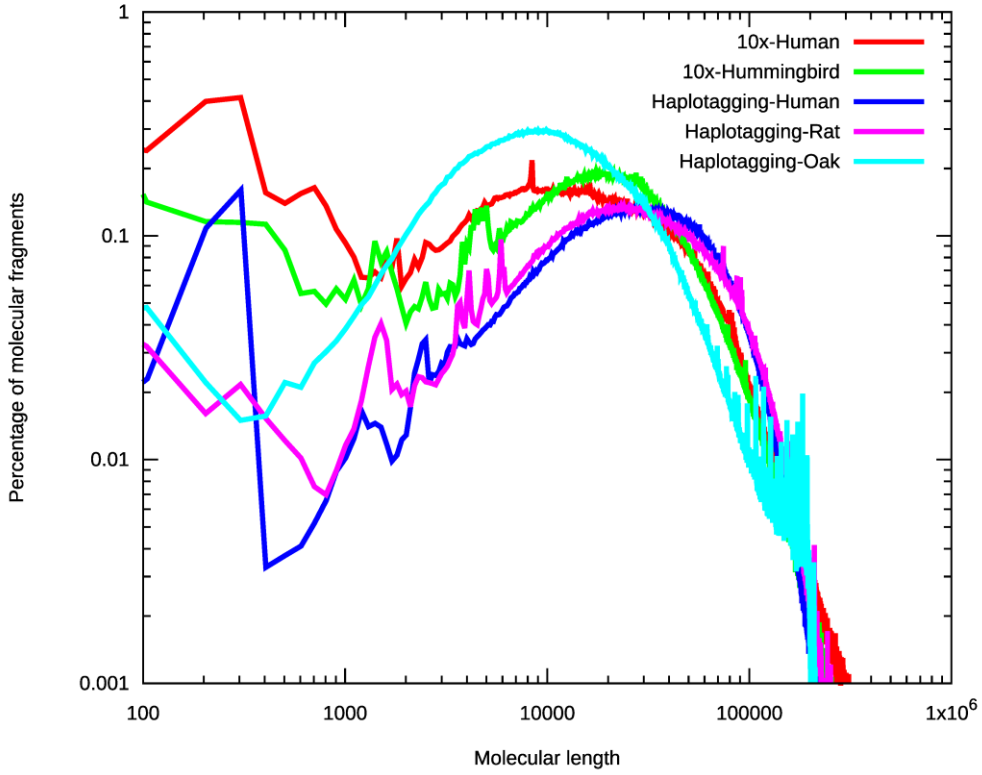
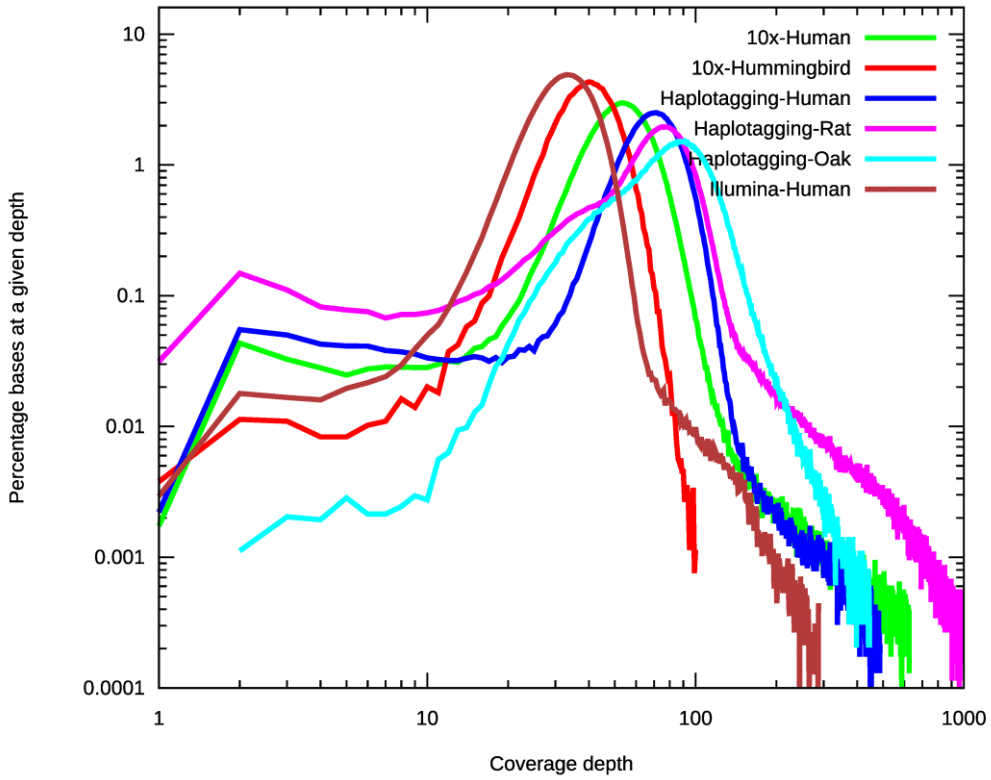

1
0.1
0.01
0.001
Percentage of molecular fragments
100
1000
10000
100000
1x10^6
Molecular length
10x-Human
10x-Hummingbird
Haplotagging-Human
Haplotagging-Rat
Haplotagging-Oak

Figure 4

A B


10x-Human
10x-Hummingbird
Haplotagging-Human
Haplotagging-Rat
Haplotagging-Oak
Illumina-Human
Percentage bases at a given depth
Coverage depth
10
1
0.1
0.01
0.001
0.0001
1
10
100
1000

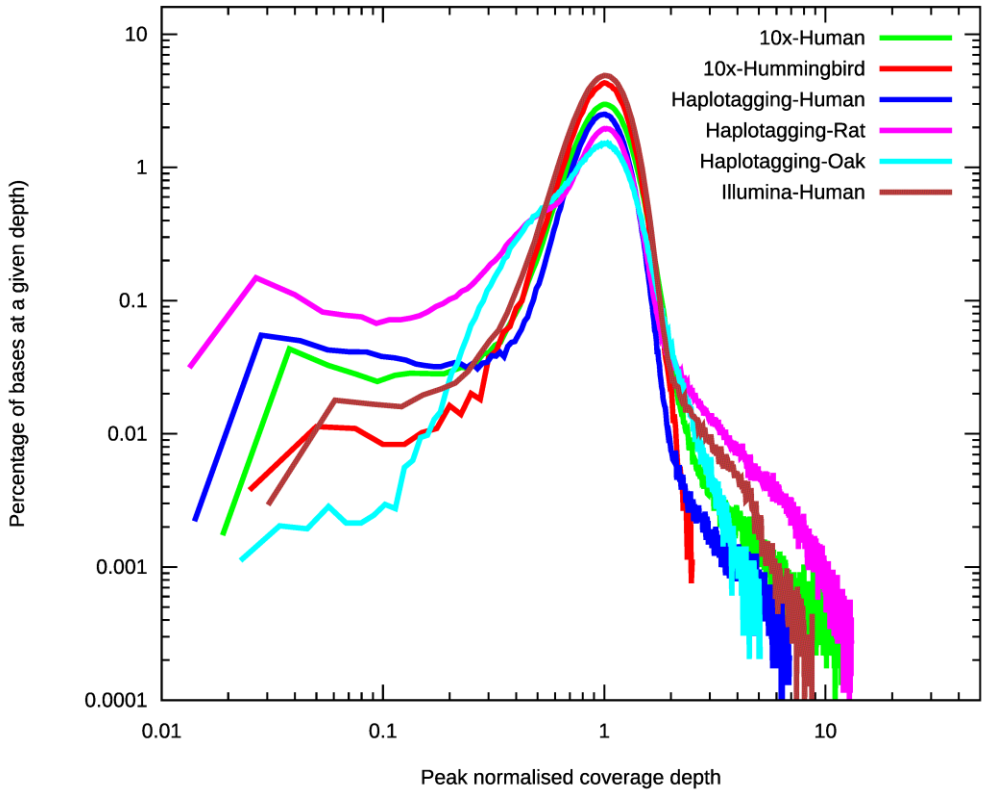

10x-Human
10x-Hummingbird
Haplotagging-Human
Haplotagging-Rat
Haplotagging-Oak
Illumina-Human
Percentage of bases at a given depth)
Peak normalised coverage depth
10
1
0.1
0.01
0.001
0.0001
0.01
0.1
1
10

Figure 5

A B C

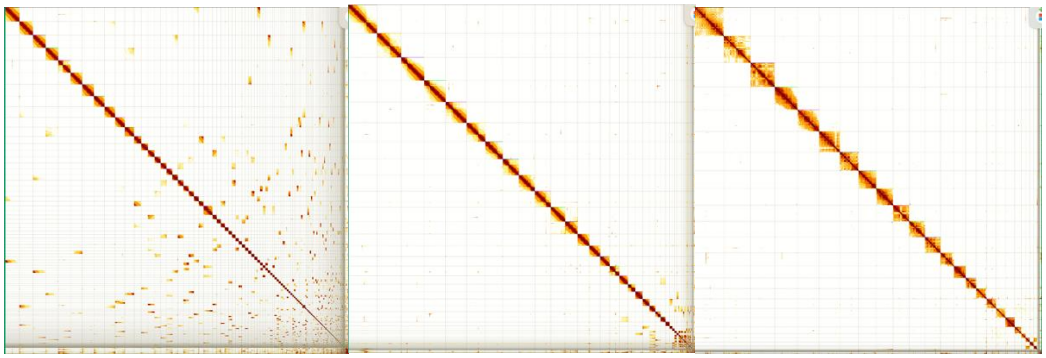


Figure 6

Table 1: Features of Hi-C reads with three different datasets

Datasets	Read pairs	Unmapped (%)	PCR duplicated (%)	Translocated (%)	Accessibility	N50 (Mb)	N20 (Mb)	N10 (Mb)
Arima V2 NA12878-CEU	352,429,304	20.9	6.8	12.7	0.596	47.9	96.3	130.3
Arima V2 NA24385-AJ	413,162,798	24.8	6.1	16.2	0.529	47.2	104.9	141.9
Arima V1 NA12878-CEU	415,173,112	28.6	10.1	18.5	0.328	28.2	63.1	100.0

Table 2: Features of 10x and Haplotagging linked reads technology with different samples. For the Tools, LR+S means the LongRanger tool and scaff10x, whilst EMA+S means EMA and scaff10x.

Species	Platform	Tools	Read pairs	Unmapped Reads (%)	PCR duplication (%)	Accesibility	Moleular length (N50)	Barcode N50 (>=5 reads)	Barcode N50 (>=3 reads)	Reads clustered (%)
Human-NA12878	10x	LR+S	669,583,370	10.1	21.0	0.689	94,611	81	80	88.2
Hummingbird	10x	LR+S	159,605,373	14.9	5.0	0.801	63,292	22	21	87
Human	Haplotagging	EMA+S	678,683,208	2.52	30.7	0.678	73,294	12	11	55
Rat	Haplotagging	EMA+S	742,824,305	2.5	30.0	0.675	78,250	11	11	50
Oak	Haplotagging	EMA+S	208,869,403	4.17	39.5	0.563	54,969	11	8	50

Table 3: Coverage Evenness statistics

	Datasets	Coverage Mean	Coverage Variance	Unevenness
Human	Arima V2 NA12878-CEU	31.2	195.0	5.3
	Arima v2 NA24385-AJ	38.8	273.9	6.1
	Arima V1 NA12878-CEU	32.1	363.1	10.3
	10x NA12878	57.2	565.6	8.9
	Haplotagging	73.0	446.0	5.1
	Illumina	35.1	131.2	2.7

Nonhuman	10x Hummingbird	41.8	98.6	1.36
	Haplotagging Rat	83.3	3715.1	43.6
	Haplotagging Oak	90.64	1133.9	11.5

Table 4: Assembly stats from different Hi-C datasets

Data and Assembly	Total Bases (Gb)	Number of Sequences	SEQUENCE LENGTH (MB)			
			Mean	N50	N90	Maximum
HIFI READS – HG002	167.76	14,949,433	0.011	0.011	0.009	0.021
CONTIGS WITH HIFIASM	2.866	1,126	2.54	45.1	6.93	116
CONTIGS AFTER PURGE_DUPS	2.83	434	6.52	45.1	8.46	116
SCAFFOLDS WITH ARIMA V1	2.83	162	17.5	152	78.4	324
SCAFFOLDS WITH ARIMA V2	2.83	152	18.6	144	75.12	235

Table 5: A list of analysis tools for long-range platform in different applications, detailing their properties and web address.

Application		*Software Tool*	*Year*	*Properties*	*URL*
Hi-C	Genome Scaffolding	LACHESIS	2013	Deterministic, agglomerative hierarchical clustering	https://github.com/shendurelab/LACHESIS
		dnaTri	2013		https://github.com/NoamKaplan/dna-triangulation
		GRAAL	2014	Probabilistic MCMC	https://github.com/koszullab/GRAAL
		instaGRAAL	2020	Probabilistic MCMC, refined for large genomes	https://github.com/koszullab/instaGRAAL
		SALSA2	2019	Novel iterative scaffolding method	https://github.com/marbl/SALSA
		3D-DNA	2017	Deterministic best-neighbour, megascaffold approach	https://github.com/aidenlab/3d-dna
		HiRise	2016	Maximum Likelihood algorithm, official Dovetail product	https://github.com/DovetailGenomics/HiRise_July2015_GR
		ALLHIC	2019	Deterministic hierarchical clustering on autopolyploid or heterozygous genomes	https://github.com/tangerzhang/ALLHiC
		HIC-Hiker	2020	Probabilistic, dynamic programming approach to improve quality of already-scaffolded data	https://github.com/ryought/hic_hiker
		EndHic	2021	Improves quality of already-scaffolded data	https://github.com/fanagislab/EndHiC
		YaHS	2022	Probabilistic, novel inference algorithm	https://github.com/c-zhou/yahs
	Variation detection	HiCnv	2018	Detects Copy Number Variations using Hidden Markov Models	https://github.com/ay-lab/HiCnv
		OneD	2018		https://github.com/qenvio/dryhic
		HiCtrans	2018	Detects Translocations using Hidden Markov Models	https://github.com/ay-lab/HiCtrans
		HiNT	2020	Detects both CNV and Translocations	https://github.com/parklab/HiNT
		HiTea	2021	Identifies mobile transposable element insertions	https://github.com/parklab/HiTea
		NeoLoopFinder	2021	Finds SV-induced chromatin loops	https://github.com/XiaoTaoWang/NeoLoopFinder
		EagleC	2022	Deep-Learning method for full-spectrum SV detection	https://github.com/XiaoTaoWang/EagleC
Chain-Linked Reads	Genome Scaffolding	fragScaf	2014	Agglomerative hierarchical clustering	https://github.com/adeylab/fragScaff
		Architect	2016		https://github.com/kuleshov/architect
		ARCS	2018	Designed specifically for 10x	https://github.com/bcgsc/arcs
		ARKS	2018	k-mer mapping for improved efficiency in ARCS	
		ARBitR	2021	Explicitly designed for multiple CLR platforms	https://github.com/markhilt/ARBitR
		SLR-superscaffolder	2021	Divisive hierarchical clustering	https://github.com/BGI-Qingdao/SLR-superscaffolder
	de novo Assembly	Supernova	2017	Official 10x assembly product	https://support.10xgenomics.com/de-novo-assembly/software/overview/latest/welcome
		cloudSPAdes	2019	De Bruijin assembler, extensible to metagenomic or hybrid data	https://github.com/ablab/spades/releases/tag/cloudspades-paper
		Ariadne	2021	cloudSPAdes module, deconvolves barcodes accurately	https://github.com/lauren-mak/ariadne
		Athena	2018	Improves metagenomic assembly	https://github.com/abishara/athena_meta
		TuringAssembler	2020	Introduced explicitly for TELL-Seq data	https://universalsequencing.com/software/
		TELL-seq data analysis pipeline	2020		
	Variation Detection	Long Ranger	2019	Official 10x variation detection tool, uses augmented GATK approach	https://support.10xgenomics.com/genome-exome/software/downloads/latest
		GROC-SVs	2017	Simialr to Long Ranger, uses local assembly to improve resolution	https://github.com/grocsvs/grocsvs
		NAIBR	2018	Probabilistic model using “split molecule” approach	https://github.com/raphael-group/NAIBR
		VALOR	2017	Detects genomic inversion from “split molecule” signature	https://github.com/BilkentCompGen/valor
		VALOR2	2020	Exppanded from VALOR to detect more types of SV	
		ZoomX	2018	Novel probabilistic approach to detect large rearrangements	https://bitbucket.org/charade/zoomx/src
		LEVIATHAN	2021	Can detect SVs in highly fragmented and heterozygosity genomes	https://github.com/morispi/LEVIATHAN

Supplementary

Evenness Metric

In the ideal case, the sampling of the genome would be perfectly uniform, such that every base was covered exactly the same number of times. Since this is practically impossible, we would instead prefer that every base had the same chance of being covered, and allowing for some statistical noise. If we model the sequencing process as one which samples each base of the genome at a mean rate λ, which is independent of the sampling rate of other bases, then the probability that a given base enters the library k times (i.e. has a coverage of k) is:

$$p(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} = \mathcal{P}(k|\lambda)$$

This is the standard Poisson distribution, and would be the result of a perfectly even sampling of the genome. If, however, there is not a single value of λ, but multiple different values, such that the probability of a given value of λ is given by the distribution function $f(\lambda)$, then the probability of finding a coverage value of k is given by a Polypoisson distribution, $p(k|f)$ such that:

$$p(k|f) = \int_0^\infty f(\lambda) \frac{\lambda^k e^{-\lambda}}{k!} \, \mathrm{d}\lambda.$$

The integral is carried out over the full support of the parameter λ, that being the half-infinite interval. We note that if f is a normalized distribution function on this interval, then the total probability still obeys:

$$\sum_{k=0}^{\infty} p(k|f) = \int_0^\infty f(\lambda) \left(\sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} \right) \mathrm{d}\lambda$$

$$= \int_0^\infty f(\lambda) \mathrm{d}\lambda = 1$$

The mean and the variance of a Polypoisson distribution are found from:

$$\langle k \rangle = \sum_{k=0}^{\infty} k \, p(k|f) = \int_0^\infty \lambda f(\lambda) \mathrm{d}\lambda$$

$$\mathrm{Var}(k) = \left(\sum_{k=0}^{\infty} k^2 \, p(k|f) \right) - \langle k \rangle^2 = \int_0^\infty \lambda^2 f(\lambda) \mathrm{d}\lambda + \langle k \rangle - \langle k \rangle^2$$

We note that the dimensional conflict of $\langle k \rangle$ and $\langle k \rangle^2$ appearing in linear combinations is not a problem since the Poisson distribution inherently only deals with dimensionless ‘counts’. Writing the results above in terms of the variance and mean of f, we find that:

$$\langle \mathrm{coverage} \rangle = \langle f \rangle$$

$$\mathrm{Var}(\mathrm{coverage}) = \mathrm{Var}(f) + \langle f \rangle$$

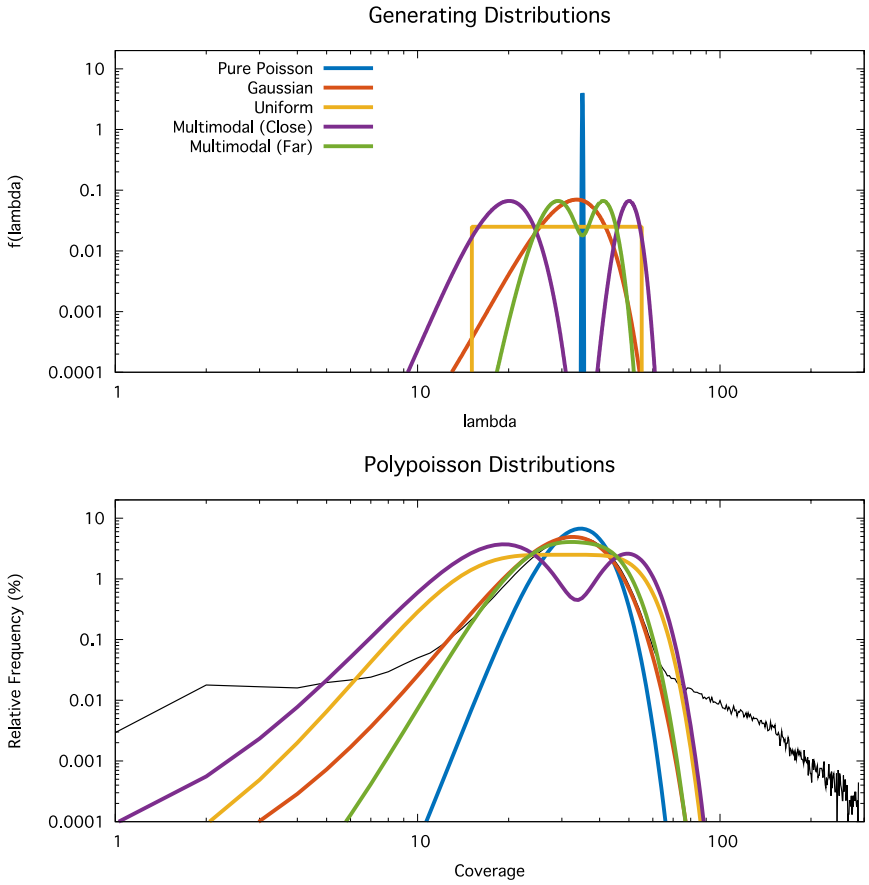
Previous works have set $f(\lambda)$ equal to the Gamma distribution, in which case $p(k|f)$ is equal to the Negative Binomial Distribution. However, we note that there is no particular need to assign a functional form to f, since all we are interested in is the dispersion of this relationship around the mean. The index of dispersion is given by:

$$D(f) = \frac{\mathrm{Var(f)}}{\langle f \rangle}$$

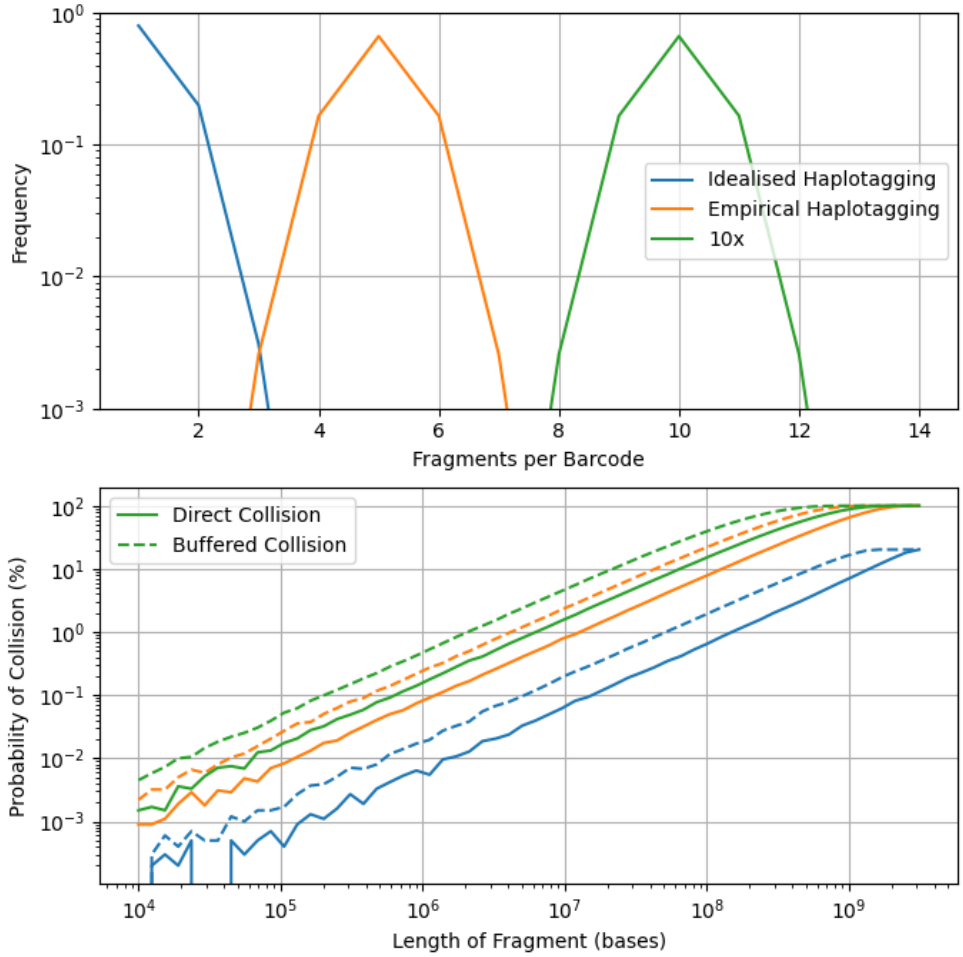
Hence:

$$D = \frac{\mathrm{Var(coverage)} - \langle \mathrm{coverage} \rangle}{\mathrm{Var(coverage)}}$$

We reiterate that this is a measure of the dispersion of f around its mean, and is hence a measure of how Poisson-like the data is: zero indicates the data is perfectly Poisson like, whilst larger values indicate that there are a significant number of processes altering how the genome is sampled. We therefore use this quantity as a metric of the unevenness of the coverage of the genome, $\mathcal{U} = D$, with smaller values being indicative of a more even coverage.



Supplementary Figure 1: An example of how generating distributions $f(\lambda)$ (top) result in different Polypoisson distributions (bottom). All distributions are chosen to have the same mean as the black curve (the Illumina human data from Fig. 3), but with other parameters chosen for demonstration purposes rather than to provide a good fit to the data. The multimodal models demonstrate that even though the Poisson distribution is monomodal, suitable generating functions can generate multimodel Polypoisson distributions.



Supplementary Figure 2: (Top) various distributions of the number of fragments of HMW-DNA which share a barcode, the blue and orange curves are designed to approximate haplotagging, whilst the purple and brown demonstrate 10x. (Bottom) the probability of barcode 'collisions' which result, as a function of the length of the fragment, assuming a diploid genome length of 6.3Gb. Solid lines demonstrate direct collisions: overlapping fragments which share the same barcode, whilst the dashed line shows 'buffered collisions', where the shared-barcode fragments do not overlap, but are too close together for reads to be unambiguously assigned to one or the other.

Supplementary Table 1: Collision-Frequency Analysis of the Chain-Linked Read platforms

Datasets	Platform	Genome Length	Fragments-per-barcode	1% Collision Size (kb)	Mean Fragment Length (kbp)	Collision Frequency (%)
Human-NA12878	10x	6.3Gb	10	2,100	59.2	0.03
Hummingbird	10x	1.8Gb	10	580	44.6	0.08
Human	Haplotagging	6.3Gb	5	4,200	56.2	0.01
Rat	Haplotagging	5.5Gb	5	3,700	57.2	0.02
Oak	Haplotagging	1.4Gb	5	970	38.5	0.04

Instructions on running assembly pipelines

Software packages

scaffHic
https://github.com/wtsi-hpag/scaffHiC
Note: scaffHiC contains PretextMap and we here use scaffHiC to process data and generate Hi-C maps as well as length distributions. We did not use it for scaffolding as yahs is noteably better in genome scaffolding.

PretextView
https://github.com/wtsi-hpag/PretextView

purge_dups
https://github.com/dfguan/purge_dups

yahs
https://github.com/c-zhou/yahs

samtools
https://github.com/samtools/

Produce sorted bam file - AJ.bam

/nfs/users/nfs_z/zn1/src/scaffHiC/src/scaff-bin/bwa-mem2 mem -t 54 -5SPM GRCH38.fasta
/lustre/scratch117/sciops/team117/hpag/zn1/project/HiC/arima/human/QC/GM24385.AJ.R1.fastq.gz
/lustre/scratch117/sciops/team117/hpag/zn1/project/HiC/arima/human/QC/GM24385.AJ.R2.fastq.gz > align-AJ.sam
samtools view -@ 50 -bS align-AJ.sam > Sorted_names.bam
samtools fixmate -@ 50 -m Sorted_names.bam Fixmate.bam > try.out
samtools sort -@ 50 -o Sorted.bam Fixmate.bam > try.out
rm -rf align-AJ.sam Sorted_names.bam Fixmate.bam
samtools markdup -@ 50 -r -s Sorted.bam Dupmarked.bam > try.out
mv Dupmarked.bam AJ.bam

Coverage analysis

samtools depth Sorted.bam | egrep _0 | awk '($2%100==0){print $0}' > depth.dat
sort -n -k 3 depth.dat | awk '{print $1,$3}' > depth-raw.dat
/nfs/users/nfs_z/zn1/src/scaffHiC/src/scaff-bin/distribution_hic-coverage depth-raw.dat | awk '{print $2,$3}' > depth-freq.dat

Hi-C contact map

/nfs/users/nfs_z/zn1/src/scaffHiC/src/scaffhic -nodes 54 -depth 50 -score 200 -map arima-AJ.map -plot arima-AJ.png -length 500000 -file 0 -fq1 /lustre/scratch117/sciops/team117/hpag/zn1/project/HiC/arima/human/QC/GM24385.AJ.R1.fastq.gz -fq2 /lustre/scratch117/sciops/team117/hpag/zn1/project/HiC/arima/human/QC/GM24385.AJ.R2.fastq.gz GRCH38.fasta aj-arima.fasta > try.out

Here we obtained arima-AJ.map and arima-AJ.png
You may use PretextView to view the Hi-C map:
https://github.com/wtsi-hpag/PretextView

Genome assembly

Contigs
~zn1/src/hifiasm/hifiasm -o hg002-hifiasm -t 80 HG002-HiFi-all.fastq.gz > try.out
egrep "^S" hg002-hifiasm.p_ctg.gfa | awk '{print ">"$2"\n"$3}' > hg002-hifiasm.fasta

Purge_dups
/nfs/users/nfs_z/zn1/src/minimap2/minimap2-2.17_x64-linux/minimap2 -t 30 -xmap-pb hg002-hifiasm.fasta HG002-HiFi-all.fastq.gz | gzip -c - > align.paf.gz
/nfs/users/nfs_z/zn1/src/purge_dups/bin/pbcstat align.paf.gz
/nfs/users/nfs_z/zn1/src/purge_dups/bin/calcuts PB.stat > cutoffs
/nfs/users/nfs_z/zn1/src/purge_dups/bin/split_fa hg002-hifiasm.fasta > Human.split
/nfs/users/nfs_z/zn1/src/minimap2/minimap2-2.17_x64-linux/minimap2 -t 20 -xasm5 -DP Human.split Human.split | gzip -c - > split.self.paf.gz
/nfs/users/nfs_z/zn1/src/purge_dups/bin/purge_dups -2 -T cutoffs -c PB.base.cov split.self.paf.gz > dups.bed
/nfs/users/nfs_z/zn1/src/purge_dups/bin/get_seqs dups.bed hg002-hifiasm.fasta > purged.fa 2> hap.fa

Scaffolding
/nfs/users/nfs_z/zn1/src/scaffHiC/src/scaff-bin/bwa-mem2 mem -t 54 -5SPM purged.fa
/lustre/scratch117/sciops/team117/hpag/zn1/project/HiC/arima/human/QC/GM24385.AJ.R1.fastq.gz
/lustre/scratch117/sciops/team117/hpag/zn1/project/HiC/arima/human/QC/GM24385.AJ.R2.fastq.gz > align-purge.sam
samtools view -@ 50 -bS align-purge.sam > Sorted_names.bam

```
samtools fixmate -@ 50 -m Sorted_names.bam Fixmate.bam > try.out
samtools sort -@ 50 -o Sorted.bam Fixmate.bam > try.out
rm -rf align-AJ.sam Sorted_names.bam Fixmate.bam
samtools markdup -@ 50 -r -s Sorted.bam Dupmarked.bam > try.out
mv Dupmarked.bam AJ-scaff.bam

~zn1/src/yahs/yahs -o HG002-yahs.fa purged.fa AJ-scaff.bam > try.out

Hi-C map for scaffolded assembly
/nfs/users/nfs_z/zn1/src/scaffHiC/src/scaffhic -nodes 54 -depth 50 -score 200 -map yahs-final-AJ.map -plot yahs-final-AJ.png -
length 500000 -file 0 -fq1
/lustre/scratch117/sciops/team117/hpag/zn1/project/HiC/arima/human/QC/jz1/GM24385.AJ.R1.fastq.gz -fq2
/lustre/scratch117/sciops/team117/hpag/zn1/project/HiC/arima/human/QC/jz1/GM24385.AJ.R2.fastq.gz HG002-yahs.fa arima-
AJ.fasta > try.out

Here we have yahs-final-AJ.map and yahs-final-AJ.png.
```