Aquila_stLFR: assembly based variant calling package for stLFR and hybrid assembly for linked-reads

Xin Zhou [1,*], Lu Zhang [2,3], Xiaodong Fang [4], Yichen Liu [5], David L. Dill [1], and Arend Sidow [2,6,*]

[1]Department of Computer Science, Stanford University, Stanford, California 94305, USA

[2]Department of Pathology, Stanford University School of Medicine, Stanford, California 94305, USA

[3]Department of Computer Science, Hong Kong Baptist University

[4]BGI Tech, BGI Genomics, Shenzhen, China

[5]School of information and communication engineering, University of Electronic Science and Technology of China, China

[6]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA

Abstract

Human diploid genome assembly enables identifying maternal and paternal genetic variations. Algorithms based on 10x linked-read sequencing have been developed for de novo assembly, variant calling and haplotyping. Another linked-read technology, single tube long fragment read (stLFR), has recently provided a low-cost single tube solution that can enable long fragment data. However, no existing software is available for human diploid assembly and variant calls. We develop Aquila_stLFR to adapt to the key characteristics of stLFR. Aquila_stLFR assembles near perfect diploid assembled contigs, and the assembly-based variant calling shows that Aquila_stLFR detects large numbers of structural variants which were not easily spanned by Illumina short-reads. Furthermore, the hybrid assembly mode Aquila_hybrid allows a hybrid assembly based on both stLFR and 10x linked-reads libraries, demonstrating that these two technologies can always be complementary to each other for assembly to improve contiguity and the variants detection, regardless of assembly quality of the library itself from single sequencing technology. The overlapped structural variants (SVs) from two independent sequencing data of the same individual, and the SVs from hybrid assemblies provide us a high-confidence profile to study them.

Availability: Source code and documentation are available on `https://github.com/maiziex/Aquila_stLFR`.

1 Introduction

Generating a precise and customized diploid human genome for each individual will be a breakthrough for uncovering the fundamental relationship between genotype and phenotype, and will

*Corresponding Author. Email: xzhou15@cs.stanford.edu

have far reaching health implications, such as cancer-related variants, and risk for genetic disease [1]. Illumina short-read sequencing has had a major influence on human genetic studies. Over 100,000 individual personal genomes have been sequenced, allowing detection of unique variations in personal genomes that cause diseases [2]. Large-scale genome studies, such as the 1000 Genome Project and the 10k UK Genome Project, have relied on reference-based assembly approaches and have made great progress in uncovering genomic differences among individuals, but identifying individual variations in highly variable or repetitive regions has been less accurate due to limitations of the resequencing technology [3]. Structural variants (SVs) are also challenging to detect by alignment-based variant calling algorithms.

De novo assembly is a better alternative for building a precise diploid genome on a large scale. It has been widely used for next-generation sequencing (NGS) data, single nucleotide polymorphisms and small variants that can be detected straightforwardly by short reads [4, 5, 6]. Assembly-based structural variants detection offers a powerful approach to identify SVs [7, 8, 9]. However, the breakpoints of large variants ($50bp$) are less likely to be spanned through short reads. Third generation sequencing data (PacBio and NanoPore) use long reads can resolve this problem, but they introduce high sequencing errors and great cost for performing whole genome sequencing (WGS) at large scale [10, 8].

Recently developed 10x linked-reads and single tube long fragment read (stLFR) sequencing technologies offer cost-effective solutions for large-scale "perfect genome" assembly [11, 12, 13]. Compared to next generation sequencing data and other third sequencing technologies, these two linked-reads methods provide both low sequencing error and long-range contiguity. The long-range information of 10x linked-reads from short-read sequencing data allows detection of structural variants, *de novo* mutations, and haplotype phasing much easier and accurately [14, 15]. stLFR enables co-barcoding of over 8 million $20 - 300kb$ genomic DNA fragments. Long-range information from stLFR enables phasing variants efficiently and results in long phase block N50 ($34MB$ for NA12878, Wang2019).

The utility of these linked-reads sequencing data in generating a diploid assembly and detecting variants from assembly require development of new algorithms. The current state-of-the-art algorithm, Supernova, was introduced by 10x Genomics to assemble 10x linked-reads sequencing data, especially for the standard library with suggested sequencing coverage (mean fragment length 40kb, optimal coverage: 37X - 56X) [16]. Its performance had limitations in both assembly quality and in identifying assembly-based variant calls [17, 18, 19]. Aquila was developed recently to solve this problem for 10x linked-reads, generating perfect diploid assembly with long contiguity and achieving varians detection from assemblies with great sensitivity and accuracy in all types of variants [19].

So far, there is no available assembly or assembly-based variants-calling algorithms for stLFR, or even a universal algorithm for both stLFR and 10x linked-reads. Here, we develop Aquila_stLFR, which extends Aquila to adapt to the key characteristics of stLFR, and further introduce a hybrid assembly mode "Aquila_hybrid" to allow assembly combining both stLFR and 10x linked-reads. Aquila_stLFR and Aquila_hybrid integrate long-range phasing information to refine reads for local assembly in small phased chunks of both haplotypes, and then concatenate them basing on a

high-confidence profile, to achieve more precise and phased contiguous sequences, diploid contigs. These diploid assembled contigs allow us to detect all types of variants through simple pairwise alignments and comparisons.

2 Methods

Aquila_stLFR is a reference-assisted local *de novo* assembly pipeline (Figure 1). The reference is used to globally allocate long fragments into genomic regions, and local assembly is performed within small phased chunks for both haplotypes. The input files for Aquila_stLFR consist of a FASTQ file with raw paired reads, a BAM file and a VCF file (by FreeBayes, [20]). To generate the BAM file through bwa-mem ([21]), each read header of FASTQ file contains the barcode sequence, starting with "BX:Z:" (for instance, "BX:Z:540_839_548" where 540_839_548 is the barcode - check details in Github). This way each read in the BAM file also includes the field "BX:Z:" for Aquila_stLFR to reconstruct long fragment reads (LFRs) (Figure S1 and S2). For LFR technology, co-barcoded reads can form one individual LFR. Aquila_stLFR reconstructs all LFRs based on this concept. However, barcode deconvolution is still necessary for some LFRs since one barcode per LFR concept is not ideally implemented in real library preparation. There is a boundary threshold to differentiate two LFRs with the same barcode when the distance between two successive reads with the same barcode is larger than 50kb.

2.1 Haplotyping algorithm for LFRs

In the first step, Aquila_stLFR applies a recursive clustering algorithm to perform haplotyping reconstructed LFRs [19]. After reconstructing all LFRs, Aquila_stLFR assigns the alleles of heterozygous SNPs to each LFR by scanning the reads belonging to each LFR and comparing to the VCF file generated by FreeBayes. At a heterozygous locus 0 is the reference allele and 1 is the alternate allele. Ideally, there should be two clusters for each pair of heterozygous SNPs: one cluster with all LFRs supporting the maternal haplotype (for an instance, "01"), and another cluster with all LFRs supporting the paternal haplotype (the complementary format,"10"). However, Aquila_stLFR could also detect two other clusters with fewer LFRs supporting the wrong haplotypes ("00" or "11") that are caused by sequencing error. Aquila_stLFR uses a Bayesian probability model to rule out the two clusters with wrong haplotypes for each pair of heterozygous SNPs [19]. After excluding all the clusters with wrong haplotypes, the two remaining clusters form the correct maternal and paternal haplotypes. Aquila_stLFR then recursively aggregates small clusters into big ones for each haplotype relying on a supporting threshold. For instance, two clusters are merged if the number of molecules supporting the same haplotype exceeds this threshold in both of them. This threshold is set to 3 by default, corresponding to a merging error percentage $\leq ((1-p_1)(1-p_2))^3$ (for each pair of variants, if each variant matched the true variant with probability $p1$ and $p2$, respectively). Aquila_stLFR performs clustering recursively until no more clusters can be merged based on the supporting threshold. To further extend the phase blocks, Aquila_stLFR similarly performs recursive clustering when two phase blocks have a number of overlapping variants greater than a pre-defined threshold. The threshold is set to 5 by default so that the merging error due to sequencing error p is $\leq p^5$. When no more phase blocks can be merged the process has converged. Eventually, The LFRs within the clusters are assigned to the maternal or paternal haplotypes of the relevant phase blocks.

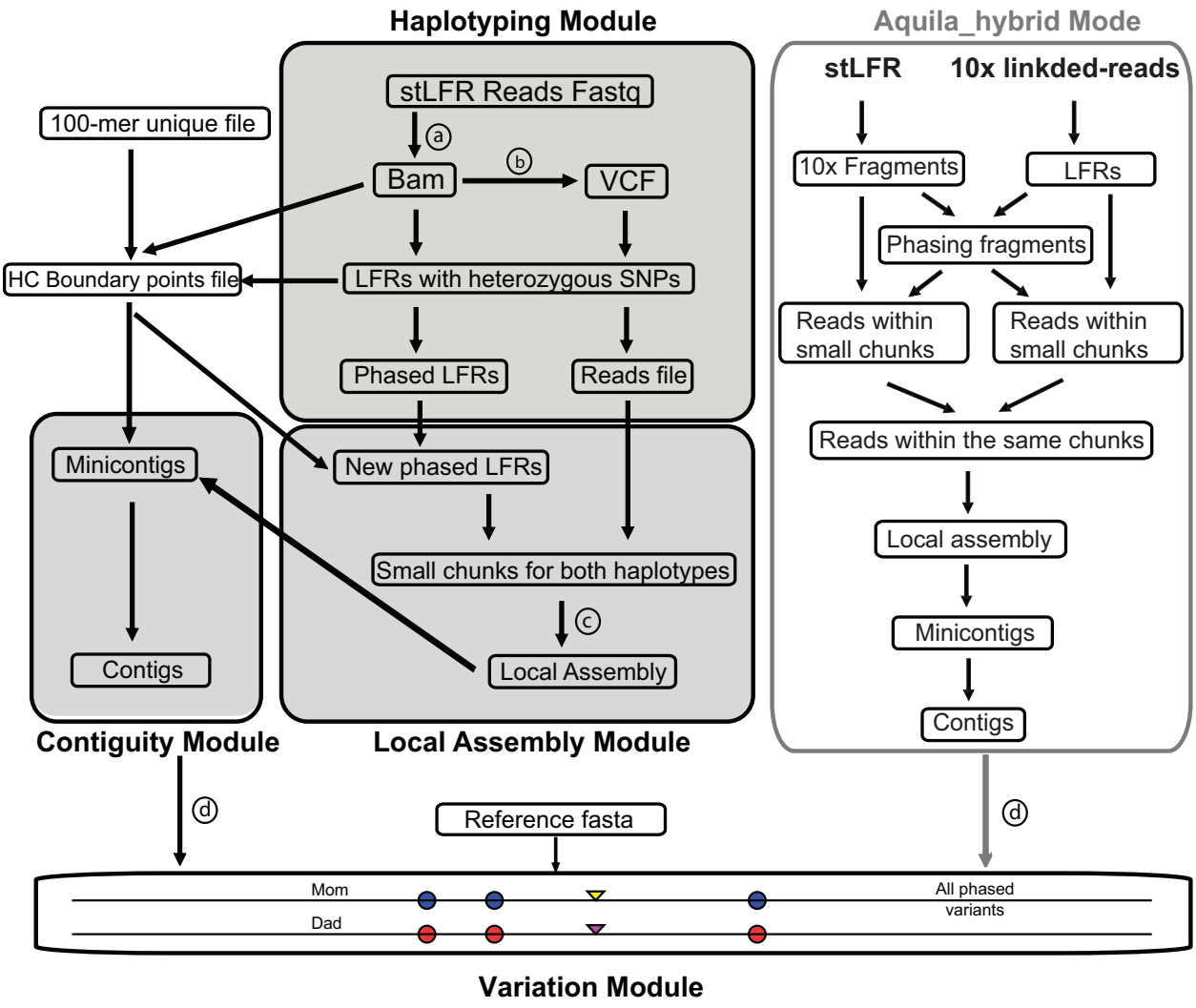

Figure 1: Pipeline of Aquila_stLFR, a reference-assisted diploid-resolved genome assembly for stLFR. Input files: FASTQ file, BAM file and VCF file. a: Bwa-mem; b: FreeBayes; c: SPAdes; d: minimap2 and paftools.

2.2 Linear local assembly

The average phase block length for the library used in this paper is approximately $282kb$, and the maximum phase block length achieved, $104Mb$. Instead of directly assembling reads within big phase blocks for both haplotypes, Aquila_stLFR turns the global assembly problem into a linear local assembly problem. It cuts large phase blocks ($\geq 200kb$ for default) into small chunks ($100kb$ by default) based on a high-confidence boundary point profile. This profile consists of human genomic regions which are covered by sufficient reads, and do not overlap with repetitive sequences. The LFRs are then allocated into small chunks for both haplotypes. Some LFRs could be allocated to more than one chunk, depending on their total length. However, Aquila_stLFR only extracts

the corresponding reads from the LFRs into each chunk based on the range of the reads within each LFR and the associated chunk. Finally, Aquila_stLFR applies local assembly for each small chunk of both haplotypes by SPAdes [22]. Local assembly within small chunks avoids ambiguous reads being assembled in a global scale, and it is not computationally intensive. The total assembly time is approximately $N_{total} * T_{onechunk}$ (N_{total}: total number of small chunks, $T_{onechunk}$: average assembly time for one small chunk). Eventually, minicontigs for both haplotypes are assembled in small chunks. To achieve large contiguity, Aquila_stLFR iteratively concatenates minicontigs into final contigs based on the high-confidence boundary point profile for both haplotypes. We use QUAST (Gurevich et al., 2013) to generate various assembly metrics such as N50 and NA50. –extensive-mis-size 1000 is applied as the lower threshold of the relocation size.

2.3 High-confidence boundary point profile

To find genomic regions not involving repetitive sequences, Aquila_stLFR re-implements the Umap from hoffman mappability [23] to allow it assemble most of diploid species, which have high quality references. It identifies the mappability of a genome for a given read length k ($k = 100$ by default). It involves three essential steps. Firstly, it generates all possible k-mers from the reference. Secondly, it uses Bowtie [24] to map these unique k-mers to the reference. Thirdly, it records the positions of all k-mers, which align to only one region in the reference.

2.4 Assembly based variants calling

Minimap2 [25] and paftools (`https://github.com/lh3/minimap2/tree/master/misc`) are integrated into Aquila_stLFR and applied to call variants from haploid assemblies. A pairwise comparison between breakpoints of variants from two haploid assemblies is then performed to detect all types of heterozygous and homozygous variants. To evaluate our diploid assembly, we generate SNPs, small indels ($< 50bp$) and SV ($\geq 50bp$). To achieve that, heterozygous variants are defined if one haploid assembly contains alternate allele(s) and the other haploid assembly contains reference allele(s). Homozygous variants are defined if both haploid assemblies contain alternate allele(s). For compound indel/SV, we split them into two heterozygous variants. Check "`--all_regions_flag=1`" for "Aquila_stLFR_assembly_based_variants_call.py" in github to perform these analyses.

2.5 Hybrid assembly for stLFR and 10x linked-reads

10x linked-reads sequencing creates millions of partitions in a microfluid system and within each partition, around 10 long DNA fragments ($10kb - 100kb$) share the same unique barcode. stLFR typically creates over 1.8 billion unique barcoded beads, and uses 10 - 50 million of them to capture 10 - 100 million long DNA fragments in a single tube. The greatest amount of sub-fragments from each long DNA fragments are co-barcoded by a unique barcode. To combine stLFR and 10x linked-reads, we introduce a "Aquila_hybrid" mode, which applies an analogous concept to reconstruct long DNA fragments and generate the same data structure for long fragments of both technologies (Figure 1). Aquila_hybrid performs efficient haplotyping for all the fragments in the next step. Based on the phasing information, if the phase block length is beyond a threshold ($200kb$), Aquila_hybrid cuts it into multiple small chunks, which is the same as that for single technology. Aquila_hybrid then extracts reads from each library independently for each phased chunk of both haplotypes, and then merges reads from the same chunks for both libraries. Linear

Sequenced Library	Sequencing Technology	Sample id	Raw coverage (X)	C_F	C_R	μ_{FL} (kb)
L1	stLFR	NA12878	98	238	0.35	25.7
L2	stLFR	NA12878	98	238	0.35	25.7
L1	10x	NA12878	103	123	0.41	79.0
L3	10x	NA12878	106	958	0.07	99.2

Table 1: Parameters of stLFR library (L1, L2) and 10x linked-read library (L1, L3) for NA12878. C_F: fragment physical coverage; C_R: read coverage per fragment. μ_{FL}: mean fragment length.

local assembly can be performed within the small phased chunks to assemble mini-contigs for both haplotypes. Finally, Aquila_hybrid achieves contiguity in a large scale based on the high-confidence boundary points profile.

2.6 Barcode "0_0_0": no specificity

By reconstructing LFRs, we detect reads with barcode "0_0_0" can span a whole chromosome which means they have no LFR specificity. We did two experiments to investigate the influence of reads with barcode "0_0_0", and found involvement of these reads would increase diploid ratio over 2 by sacrificing contiguity. Aquila_stLFR and Aquila_hybrid then remove reads with barcode "0_0_0"" for local assembly in both haplotypes.

3 Results

3.1 Characteristics of Aquila_stLFR and Aquila_hybrid assemblies

Two stLFR libraries (L1_stLFR and L2_stLFR, [26]), and two 10x linked-reads libraries (L1_10x and L3_10x, [19]) are used in this study, obtained from NA12878 (Table 1). Previous studies have shown several parameters from barcoded linked-reads libraries having influence on human diploid assembly: C_F: average physical coverage of the genome by long DNA Fragments; C_R: average coverage of short reads per fragment; $N_{F/P}$: number of fragments per partition; μ_{FL}: average unweighted DNA fragment length. The optimal physical coverage C_F is between 332X and 823X for assembly quality, the optimal length-weighted fragment length ($W_{\mu FL}$) is around 50 - 150kb, and $N_{F/P}$ has no significant influence [17].

Here, it shows the contig N50 values for both stLFR libraries are approximately $26kb$, the diploid ratio 94%, and the genome fraction 90% (Table 2). The contig contiguity is not comparable to that of 10x linked-reads libraries by Aquila ([19]) or Supernova ([16]). One limitation for stLFR libraries is that the short-reads used by stLFR are only $100bp$, which only accounts for 66.7% of short reads length ($150bp$) by 10x linked-read libraries. Another limitation is the average fragment length (μ_{FL}) is only $20kb$ for stLFR libraries. On the other side, even though stLFR sequencing technology allows much more barcodes to achieve one barcode per long fragment, the small $N_{F/P}$ does not compensate the assembly quality.

Aquila_stLFR uses the hybrid assembly mode "Aquila_hybrid", to assemble both stLFR and 10x

NA12878	Contig N50 (bp)	Contig NA50 (bp)	Diploid Fraction (%)	Genome Fraction (%)
L1_stLFR	26,682	25,723	94.0	90.77
L2_stLFR	25,566	24,658	93.8	89.51
L1_10x	34,759	31,645	98.1	95.45
L3_10x	120,963	116,438	98.7	96.17
L1_10x+L2_stLFR	44,482	40,703	98.5	91.85
L3_10x+L2_stLFR	142,234	136,274	98.8	95.39

Table 2: Assembly metrics of four libraries for NA12878. Genome Fraction, percentage of reference genome that is covered by the assembly. Diploid Fraction, percentage of Genome Fraction that is covered by exactly two parental contigs. L1_10x+L2_stLFR describes performance for a hybrid combination of the data from L1 (linked-reads) and L2 (stLFR), L3_10x+L2_stLFR describes performance for a hybrid combination of the data from L3 (linked-reads) and L2 (stLFR).

linked-read libraries. The previous study shows two 10x linked-reads libraries with good comparable assembly quality, can be combined to further improve contiguity ([19]). Here, we use the stLFR library L2_stLFR, to perform a hybrid assembly with 10x link-reads library L1_10x and L3_10x, respectively. Library L1_10x has contig N50 $35kb$, L3_10x has contig N50 $121kb$, and the combined L1+L3 (10x) can not achieve a better N50 contiguity since the assembly quality for two linked-read libraries are not comparable ([19]). Here, our hybrid assembly results show the contig N50 for hybrid library substantially increase by 74.0% - 456.3% and 17.6% - 28.0% , compared to that of stLFR, and 10x linked-reads, respectively (Table 2). It indicates that stLFR and 10x linked-reads can alway be used in a complementary fashion for assembly to improve contiguity, regardless of the assembly quality of the single library itself.

3.2 Assembly-based detection of SNPs and small indels

For assembly-based variant calling, 94% diploid ratio guarantees the correct zygosities of the variants, and allow us to detect variants in diploid assemblies (see Methods). The total numbers of assembly-based SNP calls are 3,882,250 (L1_stLFR) and 3,882,323 (L2_stLFR), compared to the total numbers of reference-based calls of 3,860,161 (L1_stLFR) and 3,860,124 (L1_stLFR). Numbers of heterozygotes or homozygotes are also comparable between the two approaches (Table S1). Compared to the GIAB SNPs callset v3.3.2 ([27]), The recall and precision of assembly-based SNPs is around 94% and 91%, respectively (Table 3). Compared to the GIAB small indel callset v3.3.2, Aquila_stLFR produces considerably more calls (e.g., 887,320 in L2 vs GiaB's 531,228; Table 3 and S2). The size distribution of Aquila_stLFR's small indels matches the one from GIAB very closely, exhibiting the same $2bp$ periodicity such that insertions or deletions of an even length are more common than those that are one base longer or shorter (Figure S3). The recall of small indels for both stLFR libraries is approximately 90%, and the precision is approximately 94% (Table 3). Furthermore, we can see that the performances of hybrid assemblies for both SNP and indels are increased comparing to that of single stLFR/10x library. This is consistent with the increased assembly contiguity by hybrid assemblies.

SNP		True Positives	False Negatives	False Positives	Genotype Mismatches	Total number	Precision	Recall	F1
L1_stLFR	**Aquila**	2,851,008	191,775	292,017	22,232	3,882,250	0.907189	0.936974	0.921841
	FreeBayes	3,019,928	22,855	35,583	4,607	4,158,048	0.988362	0.992489	0.990421
L2_stLFR	**Aquila**	2,847,211	195,572	297,034	22,339	3,882,323	0.905631	0.935726	0.920433
	FreeBayes	3,019,928	22,855	35,568	4,612	4,157,982	0.988367	0.992489	0.990424
L1_10x+L2_stLFR	**Aquila**	2,889,402	153,375	108,623	7,126	3,788,821	0.963807	0.949594	0.956648
L3_10x+L2_stLFR	**Aquila**	3,006,002	36,781	103,758	5,000	3,931,076	0.966673	0.987912	0.977177

INDEL		True Positives	False Negatives	False Positives	Genotype Mismatches	Total number	Precision	Recall	F1
L1_stLFR	**Aquila**	478,057	53,325	31,164	26,003	886,681	0.938801	0.899648	0.918808
	FreeBayes	427,285	72,413	30,666	25,904	829,341	0.934220	0.855086	0.892903
L2_stLFR	**Aquila**	477,939	53,443	31,898	38,383	887,320	0.937435	0.899426	0.918037
	FreeBayes	427,283	72,415	30,669	25,906	829,342	0.934213	0.855082	0.892898
L1_10x+L2_stLFR	**Aquila**	478,305	53,077	41,560	27,313	935,953	0.920056	0.900115	0.909976
L3_10x+L2_stLFR	**Aquila**	500,104	31,278	28,148	9,743	945,772	0.946715	0.941138	0.943918

Table 3: Comparison of SNP/indel of different pipelines with Aquila_stLFR. Variants were called from four different assemblies and compared to GIAB NISTv3.3.2. Variant counts and performance scores were generated by RTGtools/hap.py an Illumina haplotype comparison/benchmarking tool.

Our benchmarks show the assembly-based small indels calling by Aquila_stLFRR outperforms alignment-based algorithm (eg. FreeBayes), even though assembly-based SNPs calling is not close to perfect like FreeBayes. To compensate SNPs calling, Aquila_stLFR indeed integrates all the SNP calls from alignment-based algorithms (eg. FreeBayes) in the variant calling module, which are missed by assemblies.

3.3 Assembly-based detection of structural variants

From diploid assemblies, we detect over 28,000 SVs ($\geq 50bp$) in both libraries (eg. 25,837 deletions and 3,102 insertions for L1_stLFR, Table 4), and the size distribution of them indicate that stLFR assembles achieve a wide range of SVs (Figure S4).

SV calls from two independent sequencing technologies could generate high-confidence SVs. We compare the SV calls among stLFR, 10x linked-reads, and hybrid assemblies. The overlapped SVs between two libraries are defined if their breakpoints, reference and alternate allle(s) are exactly the same (no soft threshold). Our comparison results show that 27% SVs from stLFR and 33% SVs from 10x linked-reads are overlapped, which give us high confidence that these 7,809 overlapped SV calls are true positives (Table 5). We could also see that the SV calls by hybrid assembly have a larger overlap with that of 10x assembly than stLFR assembly. We also see similar trend for small indels (Table S3).

In general, We also note that the fraction of variants that are heterozygous varies over a narrow range across all types and sizes of detected variation (Figure S5), revealing no obvious biases.

Deletion	Homo	Hetero	Total	Insertion	Homo	Hetero	Total
L1_stLFR	2,715	23,122	25,837	**L1_stLFR**	436	2,666	3,102
L2_stLFR	2,675	23,516	26,191	**L2_stLFR**	431	2,689	3,120
L1_10x+L2_stLFR	2,602	13,233	15,835	**L1_10x+L2_stLFR**	616	13,940	14,556
L3_10x+L2_stLFR	3,554	13,587	17,141	**L3_10x+L2_stLFR**	808	4,892	5,700

Table 4: Different types of SVs ($\geq$ 50bp) for four assemblies. Deletions: homozygous ones, heterozygous ones and total number. Insertions: homozygous ones,heterozygous ones and total number.

	DEL($\geq$50)			INS($\geq$50)		
	Total	**Overlap**	**Unique**	**Total**	**Overlap**	**Unique**
L2_stLFR	26,191	6,347	19,844	3,120	1,462	1,658
L3_10x	17,806		11,459	6,057		4,595
L3_10x	17,806	13,452	4,354	6,057	4,391	1,666
L3_10x+L2_stLFR	17,141		3,689	5,700		1,309
L2_stLFR	26,191	6,465	19,726	3,120	1,527	1,593
L3_10x+L2_stLFR	17,141		10,676	5,700		4,173

Table 5: Overlapped and unique number of SVs ($\geq$50bp) in all regions between L2_stLFR, L3_10x, and L3_10x+L2_stLFR. For overlapped SVs between two libraries, the break points and reference and alternate alleles are the same.

4 Discussion

stLFR sequencing technology provides long range information through barcoded reads clouds. To take advantage of this long range information, Aquila_stLFR globally performs haplotyping long LFRs into two haplotypes, and then allocates short reads into small phased chunks to do local assembly. The key concept of Aquila_stLFR is that it guarantees a complete diploid assembly, which allows us to further detect all types of variants, especially for small indels and large SVs in diploid assemblies. For barcoded linked-reads technology, parameters like C_F, C_R, μ_{FL} are essential for assembly quality. 10x Genomics recommends the standard library (mean fragment length $\mu_{FL} \sim$ 40kb, optimal coverage: 37X - 56X), to assemble 10x linked-reads sequencing data ([16]). Recent studies with different customized link-reads libraries show that the optimal physical coverage C_F is between 332X and 823X and assembly quality could further improve by even higher C_F if the corresponding C_R is increased. They also suggest that the optimal length-weighted fragment length ($W_{\mu FL}$) is around 50 - 150kb ([17],[18]). We find all stLFR libraries have a similar configuration of these three parameters ($C_F = 238X$, $C_R = 0.35$, $\mu_{FL} = 25.7kb$). This indicates that the low μ_{FL} is one key factor causing the lower contiguity of stLFR assemblies comparing with that of 10x linked-reads assemblies. Furthermore, stLFR libraries use $100bp$ paired short reads which cause limitation for local assembly compared to $150bp$ paired short reads used by 10x link-reads. We believe this study can provide a guideline for future stLFR libraries preparation to achieve significant improvement in assembly and the downstream analysis.

Beyond the alignment-based variant detection, assembly-based variant detection provides us an indispensable alternative to study all types of variants, especially small indels and large SVs. To-

day, different types of sequencing technologies allow us to study variations for the same individual independently, and the further combination of these different sequencing data gives us power to improve our work qualities. Aquila_stLFR can detect a large range of SVs from stLFR libraries, and its hybrid assembly mode can efficiently assemble both stLFR and 10x link-reads sequencing data. The overlapped SVs between these two linked-reads technologies, and the SVs from their hybrid assemblies provide us a high-confidence profile to study SVs.

5 Acknowledgements

This research was supported by the Joint Initiative for Metrology in Biology (JIMB; National Institute of Standards and Technology).

References

[1] Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase information for human genomics (2011).

[2] Lunshof, J. e. a. Personal genomes in progress: from the human genome project to the personal genome project. *Dialogues Clin Neurosci* **12**, 47–60 (2010).

[3] Sohn, J. I. & Nam, J. W. The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics* (2018).

[4] Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* (2011).

[5] Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* (2015).

[6] Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome (2015).

[7] Wala, J. A. *et al.* SvABA: Genome-wide detection of structural variants and indels by local assembly. *Genome Research* (2018).

[8] Fan, X., Chaisson, M., Nakhleh, L. & Chen, K. HySA: A hybrid structural variant assembly approach using next-generation and single-molecule sequencing technologies. *Genome Research* (2017).

[9] Nattestad, M. & Schatz, M. C. Assemblytics: A web analytics tool for the detection of variants from an assembly. *Bioinformatics* (2016).

[10] Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications (2015).

[11] Peters, B., Liu, J. & Drmanac, R. Co-barcoded sequence reads from long dna fragments: a cost-effective solution for perfect genome sequencing. *Front Genet* **5**, 466 (2014).

[12] Zheng, G. e. a. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* **34**, 303–311 (2016).

[13] McElwain, M. A. e. a. Long fragment read (lfr) technology: Cost-effective, high-quality genome-wide molecular haplotyping. *Methods Mol. Biol.* **1551**, 191–205 (2017).

[14] Zheng, G. X. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology* (2016).

[15] Zhou, X., Batzoglou, S., Sidow, A. & Zhang, L. HAPDeNovo: A haplotype-based approach for filtering and phasing de novo mutations in linked read sequencing data. *BMC Genomics* (2018).

[16] Weisenfeld, I. e. a. Direct determination of diploid genome sequences. *Genome research* **5**, 757–767 (2017).

[17] Zhang, L., Zhou, X., Weng, Z. & Sidow, A. Assessment of human diploid genome assembly with 10x Linked-Reads data (2019).

[18] Zhang, L., Zhou, X., ziming Weng & Sidow, A. De novo diploid genome assembly for genome-wide structural variant detection. *bioRxiv* (2019).

[19] Zhou, X. e. a. Aquila: diploid personal genome assembly and comprehensive variant detection based on linked reads (2019). URL `https://www.biorxiv.org/content/biorxiv/early/2019/06/05/660605.full.pdf`.

[20] Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing (2012). URL `https://arxiv.org/abs/1207.3907`.

[21] Li, H. & Durbin, R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

[22] Bankevich, A. e. a. Spades: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19**, 455–477 (2012).

[23] Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: Quantifying genome and methylome mappability. *Nucleic Acids Research* (2018).

[24] Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* (2012).

[25] Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

[26] Wang, O. e. a. Efficient and unique cobarcoding of second-generation sequencing reads from long dna molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome research* **5**, 798–808 (2019).

[27] Zook, J. e. a. Reproducible integration of multiple sequencing datasets to form high-confidence snp, indel, and reference calls for five human genome reference materials (2018). URL `https://www.biorxiv.org/content/10.1101/281006v2.full`.

Appendix

Data Availability:

L1: ftp://ftp.cngb.org/pub/CNSA/CNP0000066/CNS0007597/CNX0005843/CNR0006054/

L2: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/stLFR

```
@CL100066606L2C016R059_503217   BX:Z:540_839_548
AAAACACACTTTTATTTTATTTTATTTTATTTATTTAATTTAATTTAATTTAATTTTATTTTATTTTATTTATTTTATTTTATTTTATTTTATTTTTTTG
+
A@;,<EEEE<6B=EEDAFEEEC@FECE8DE@BB=ECCCAFBA-E?ABD>BEF3EE=@332E@EB>E.EABE:D=EDAA6ECFEBE=AEEFFB@EBE?>E9
@CL100066606L2C016R059_503217   BX:Z:540_839_548
CCTGTAGTCCCAGCTACTCAGGAGGCTGAGGCAGGAGAATGGCGTGAACCCAGGAGACAGAGCTTGCAGTGAGCTGAGATAGCGTCACTGCACTCCAGCC
+
EEFFE<FEDEFFFEFEEEF?EEAF8FFD:?EFEFEAD2D6<EFEFF@FFEEFED?EAEDF3DFFFFECFFE,EFFFAFEFDCFEFFDCFFCDFFFD=<<A
@CL100066606L2C016R091_430379   BX:Z:540_839_548
CTCACCCCTTCTACCATGTGAGGACGCAGCAAGAAGTCACCTTCTATAAATCAGGAAGAGATCGCTAACCATCCCCTGTATCTGCTAGTGCCTGGATCTT
+
F5F>EFFFD*FA<FED;F>DCFF==FFDEFFFFA=F8FFFF:EFAE4E-0<7<?FA<FBF'@AFF3B/=FA@FFF<=F(ECF8FF5FFAFE8@FF?DF-=
@CL100066606L2C016R091_430379   BX:Z:540_839_548
CATATAACCTCACCCACATAGAGAAATAAAATAACTAGGATCCAGATGTTCCTGCATAAACCCTCTTGATTCTTACCCATTCACATTCTCCAGAAATAAC
+
CFDEAB6FDFEEFDE<F>4AE4=+4@FBD;:BE>EE;EE29AF:F1DF5/EE8'E;6<5%CEECBB.F0>3EBBFB;F>A3E?F;C.8<BF2&&@9;E,E
@CL100066606L2C017R026_292022   BX:Z:540_839_548
CCTTCACTTTCAGGTTTTAATGTATTCAACTCATTTGGATAAATACCAAGGAGCACAGCCTCTTCTTTTTTATTATTGAGCAATCTCAAATGCATCTCTG
+
FEEEFF.EF<FFFDEEF@E@@EFFCBFD?EEECEDAFBFCD@FEFF?F:E>?FEFD>3EFCCE455FDCE?F9DEE>D8:EF8DEFEEFEEFF=6FEFDE
@CL100066606L2C017R026_292022   BX:Z:540_839_548
CCCTCCTTTGGGAGATGCTGCTCTTGAAGTAGAAGAAAAGTGCTAACTGGAAATTCTTTGAAACCTCACTTTCTCATTGTCAGCTGGTCAAAAGAAAACA
+
FFFFF>EFFFFDFFFFFFEFFFFFEFEE:EFFFFFCFFDFFFFFF@FFFEC56EFFFDDF>9=FFFFFFFE@<FF'E>FBFCFFDGFE,F/@@E=D4EFF
```

Figure S1: A screen shot of stLFR fastq reads, it shows three pairs of short reads. Before performing reads alignment by bwa-mem, add barcode "BX:Z:barcode" at the header of each read.

```
CL200047468L1C006R038_39389     147     chr1    9997    0       100M    =       10013   -84     CAGATAACCCTAACCCTAACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC     9CFEFGFAFG7FGFEFEGFDFFG6E@FFG?FFFFGDGFFFFFFFFFGEFFFFFGFEFFGG@FF@GGFFFFGGF
FF?FGDFFGGFFFFFEGFG;FFFA<FF     NM:i:1  MD:Z:1C98       AS:i:98 XS:i:96 BX:Z:586_902_1136       RG:Z:12878:LibraryNotSpecified:1:unknown_
fc:0-43F6A2D2
CL200047469L1C011R060_87815     147     chr1    9997    0       100M    =       10013   -84     CAGATAACCCTAACCCTAACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC     FFDFFFCGFGEFGFFFFG1FFGFCFFFFGFGFFGE<FFFFFFGFFFGFGFFFEGG>FFFGFFFFFFFFFFFF
FFFEFF>FFGFFFFFEFFFFFFFFFF     NM:i:1  MD:Z:1C98       AS:i:98 XS:i:96 BX:Z:353_1476_263       RG:Z:12878:LibraryNotSpecified:1:unknown_
fc:0-64378858
CL100066606L2C006R022_551816    147     chr1    9997    0       100M    =       10007   -90     CAGATAACCCTAACCCTAACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC     FFFECDDFFFFFFDEFE>FECFFE6FEDFFGFEFF>FDBFFCFFFDFFFFDFFFFFFFFE6FFFFFFFCEFFF
FDEFFFFFFFFFFFDGFFEFFFFFFEC     NM:i:1  MD:Z:1C98       AS:i:98 XS:i:96 BX:Z:0_0_0      RG:Z:12878:LibraryNotSpecified:1:unknown_fc:0-43F
6A2D2-5F69628
CL200047469L1C009R005_5964      163     chr1    9998    0       100M    =       10005   108     CCATAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAGCCCTAACCCTAACCCT     5E&F=?.)<@E1BD2DCF7*76FE<2<.E)6<-54<)(4?<-7D75F5BEED>>8=BEEF83;8?E4A0C-AA
*8=E1A,ECB%:1AFFC7F@BC03EDB     NM:i:2  MD:Z:1G81A16     AS:i:93 XS:i:92 BX:Z:46_1232_9  RG:Z:12878:LibraryNotSpecified:1:unknown_fc:0-515
43204
CL100066606L2C017R018_461042    99      chr1    9998    0       100M    =       10044   146     CCATAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCT     >:)GG@EFFGFD0F?EGFFGGEFFGFGBFFCGFBGCFFFFFEEC<GFB=FFFFFF7DFGGFDBFGDFDGEFFD
GGBE@FFEDF5:FGFG@BFGB7EFGFF     NM:i:1  MD:Z:1G98       AS:i:98 XS:i:97 BX:Z:584_924_493        RG:Z:12878:LibraryNotSpecified:1:unknown_
fc:0-64717E9D
```

Figure S2: A screen shot of five reads from BAM file. Each read contains the barcode field "BX:Z:barcode".

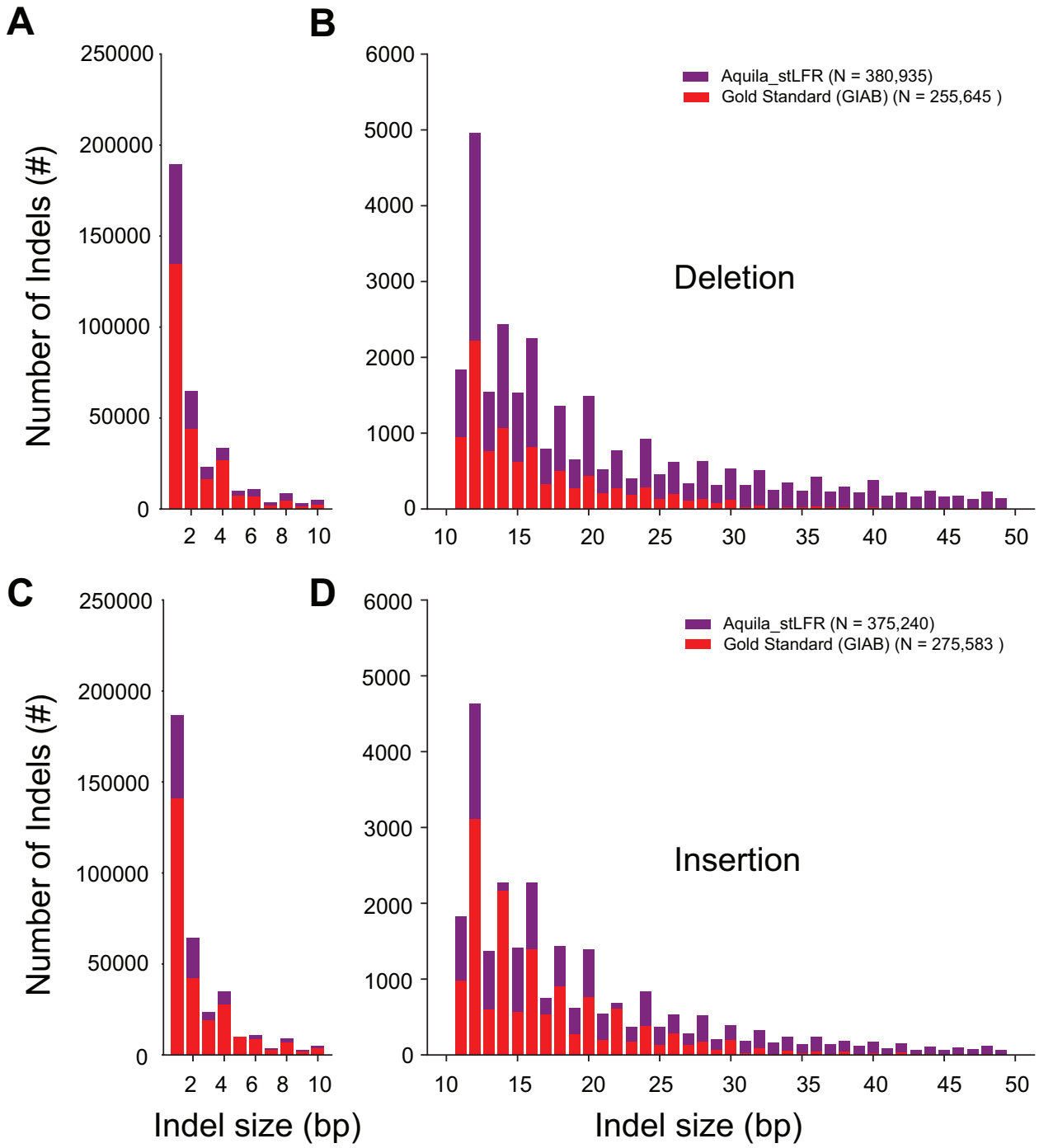

Figure S3: Small indel size distribution of NA12878 for L2 (purple plus red), with the distribution for GiaB benchmark in red only for comparison. A, deletions (≤10bp); B, deletions (>10bp and <50bp); C, insertions (≤10bp); D, insertions (>10bp and <50bp).

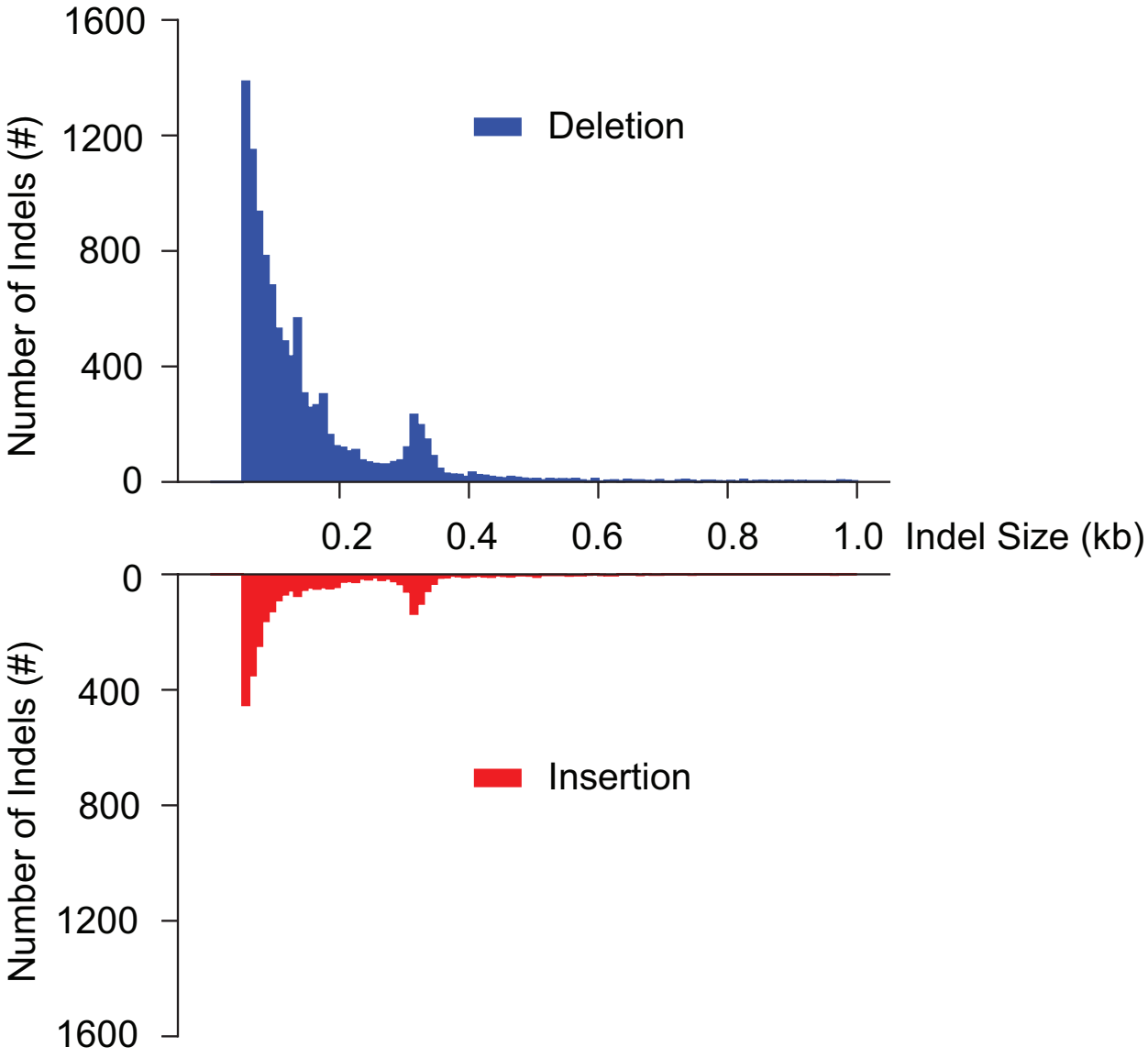

Figure S4: Histogram of SVs (≥50bp) distribution of NA12878 for L2 (only display SVs in window 50bp - 1kb).

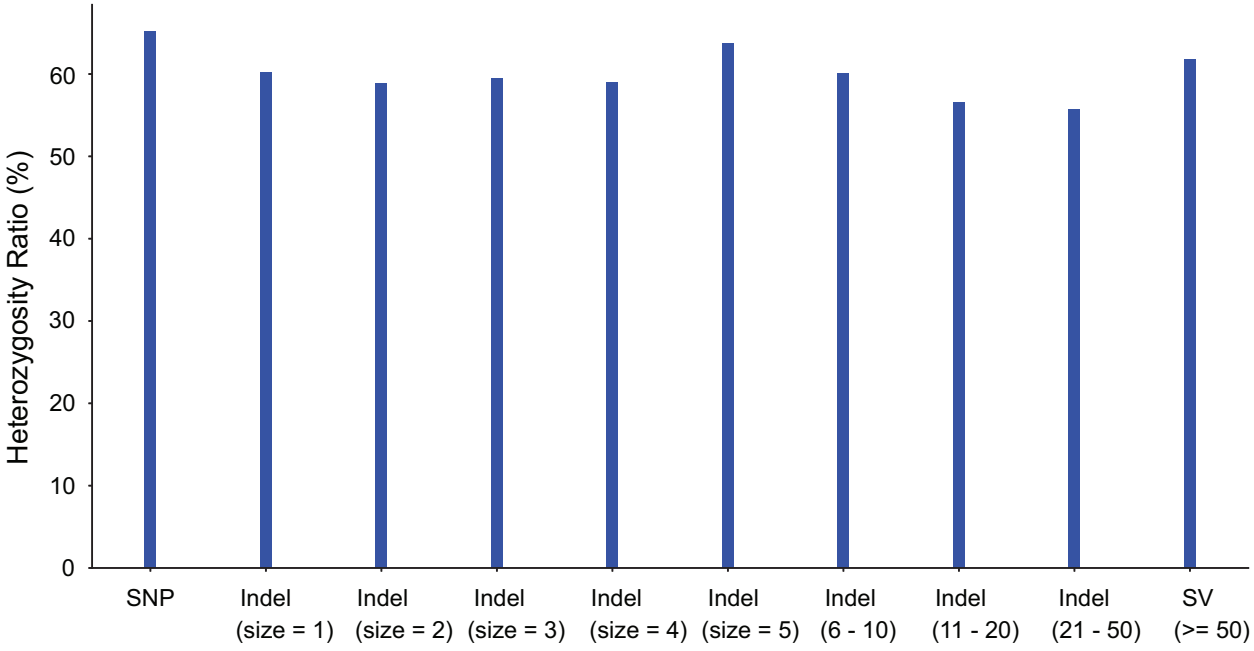

Figure S5: Average heterozygosity of all types and sizes of discovered variants of L2. It notes that the fraction of variants that are heterozygous varies over a narrow range across all types and sizes of detected variation, revealing no obvious biases.

Library	Homo Aquila_stLFR	Homo FreeBayes	Hetero Aquila_stLFR	Hetero FreeBayes	Total Aquila_stLFR	Total FreeBayes
L1_stLFR	1,321,799	1,477,019	2,560,457	2,383,142	3,882,256	3,860,161
L2_stLFR	1,317,395	1,476,985	2,564,934	2,383,139	3,882,329	3,860,124
L1_10x+L2_stLFR	1,456,867	1,477,019	2,331,971	2,383,142	3,788,838	3,860,161
L3_10x+L2_stLFR	1,468,706	1,476,985	2,462,376	2,383,139	3,931,082	3,860,124

Table S1: Comparison of the number of SNPs calls for four assemblies, using pairwise contig-to-reference alignment by Aquila_stLFR versus FreeBayes calls. Homo = homozygous, Hetero = heterozygous.

	Deletions			Insertions		
Library	**Homo**	**Hetero**	**Total**	**Homo**	**Hetero**	**Total**
L1_stLFR	110,290	376,093	486,383	107,046	316,339	423,385
L2_stLFR	110,052	376,641	486,693	106,773	316,476	423,249
L1_10x+L2_stLFR	127,910	364,455	492,365	125,025	341,881	466,906
L3_10x+L2_stLFR	139,663	368,574	508,237	136,709	338,351	475,060

Table S2: Number of Aquila_stLFR assembly-based small indel calls (<50bp) for four assemblies. Deletions: homozygous ones, heterozygous ones and total number. Insertions: homozygous ones, heterozygous ones and total number.

	DEL(<50)			INS(<50)		
	Total	**Overlap**	**Unique**	**Total**	**Overlap**	**Unique**
L2_stLFR	486,693	412,015	74,678	423,249	351,400	71,849
L3_10x	517,343		105,328	489,970		138,570
L3_10x	517,343	481,951	35,392	489,970	439,073	50,897
L3_10x+L2_stLFR	508,237		26,286	475,060		35,987
L2_stLFR	486,693	417,573	69,120	423,249	358,640	64,609
L3_10x+L2_stLFR	508,237		90,664	475,060		116,420

Table S3: Overlapped and unique number of small indels (<50bp) among L2_stLFR, L3_10x, and L3_10x+L2_stLFR. For overlapped small indels between two libraries, the break points and reference and alternate alleles are the same.