Aquila: diploid personal genome assembly and comprehensive variant detection based on linked reads

Xin Zhou[1], Lu Zhang[2], Ziming Weng[2], David Dill[1], Arend Sidow[2,3,*]

[1]Department of Computer Science, Stanford University

[2]Department of Pathology, Stanford University

[3]Department of Genetics, Stanford University

*Correspondence and requests for materials should be addressed to Arend Sidow (email: arend@stanford.edu).

Abstract

Variant discovery in personal, whole genome sequence data is critical for uncovering the genetic contributions to health and disease. We introduce a new approach, Aquila, that uses linked-read data for generating a high quality diploid genome assembly, from which it then comprehensively detects and phases personal genetic variation. Aquila achieves contiguity for both haplotypes on a genome-wide scale, and its phasing nature guarantees a real haplotype-resolved assembly instead of a haploid consensus assembly. Over 98% of a human Aquila-assembled genome is diploid, facilitating detection of the most prevalent types of human genetic variation, including single nucleotide polymorphisms (SNPs), small insertions and deletions (small indels), and structural variants (SVs), in all but the most difficult regions. All heterozygous variants are phased in blocks that can approach arm-level length. The final output of Aquila is a diploid and phased personal genome sequence, and a phased VCF file that also contains homozygous and a few unphased heterozygous variants. Aquila represents a cost-effective evolution of whole-genome reconstruction that can be applied to cohorts for variation discovery or association studies, or to single individuals with rare phenotypes that could be caused by SVs or compound heterozygosity.

Introduction

Despite recent advances, quantifying the contribution of genetic variation to specific disease risk is a stubborn biomedical problem that remains far from solved. In general, understanding the relationship between genotype and phenotype requires complete ascertainment of genotype, which for humans has yet to be achieved in a scalable fashion. At this stage in technology development, DNA sequencing still faces a vexing tradeoff between cost and completeness so that discovery of variation in larger cohorts is limited to SNPs and small indels. In fact, the relatively low cost of Illumina-based short-fragment whole genome sequencing and the even lower cost of exomes and genotyping arrays has caused considerable ascertainment bias such that the vast majority of genotype-phenotype associations focus on SNPs with small effect, even though the undetected larger variation is known to involve roughly as many bases in our genomes as SNPs and is therefore predicted to have significant phenotypic impact as well (Redon et al., 2006; Weischenfeldt et al., 2013). Also generally missing is the phasing of genetic variation, which is similarly important for estimation of phenotypic impact, as the distinction between cis and trans compound heterozygotes in an essential locus can mean the difference between health and disease (Lupski et al., 2010) and is likely to modulate risk of multigenic disease as well (Jiang et al., 2018).

Single-molecule sequencing approaches, particularly Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), provide potential solutions, as long-range information allows accurate detection of SVs and phasing (Cretu Stancu et al., 2017; Huddleston et al., 2017; Coster et al., 2018). However, the drawback of both approaches is that they exhibit poor base-pair level accuracy, leading to high error rates for SNPs and imprecise breakpoint estimation for small indels and SVs. A widely applied solution has been to supplement long reads with higher quality short read data, but these ensemble approaches are not scalable to larger cohorts due to the complexity of data generation, integration, and analysis, and have therefore been limited to small sample sizes in proof-of-principle studies (Rhoads and Au, 2015; Fan et al., 2017). A potential solution to making long reads more accurate is to sequence the same single molecule multiple times to reduce error. This is implemented in the PacBio circular consensus sequencing (CCS) approach (Travers et al., 2010; Larsen et al., 2014). However, whole human genome reconstruction with CCS can only be achieved by sequencing extremely deeply, a currently prohibitively expensive proposition for anything but small sample sizes.

A relatively recent addition to the DNA sequencing ecosystem has been pioneered by 10X Genomics, wherein the original large molecules of a gentle DNA preparation are partitioned into microfluidic compartments (Zheng et al., 2016a; Wang et al., 2019). Via a series of within-compartment molecular biology and subsequent standard steps of library construction and sequencing, barcoded short reads are produced that retain the long-range information of the long fragments of the initial DNA extract. Due to the combination of high base pair-level sequence accuracy and long-range information, 10x/Illumina data therefore support excellent SNP and small indel detection and phasing (Zheng et al., 2016a), as well as breakpoint detection of large events in cancer (Zheng et al., 2016b; Spies et al., 2017). For diploid genome reconstruction, 10x developed the *de novo* assembler, Supernova, which has been shown to produce whole human genome assemblies from 56-fold coverage 10x/Illumina data (Weisenfeld et al., 2017).

The application of assembly approaches to human genomes has been limited even though they allow powerful identification of SVs (Nattestad and Schatz, 2016, n.d.; Fan et al., 2017; Wala et al., 2018). Long-read based assemblies, such as those from PacBio data performed by FALCON-Unzip (Chin et al., 2016), exhibit respectable contiguity and variant detection but still suffer from extremely high cost (Rhoads and Au, 2015). Supernova assemblies based on 10x/Illumina data are less expensive and allow detection of all types of variation but power is limited because a substantial fraction of the genome is not assembled in a diploid state and genotyping error is still high (Zhang et al., 2019). Overall, cost-effective assembly-based approaches still suffer from incomplete resolution of the diploid genome and limited power of variant detection in a personal genome. On the other hand, assembly-based approaches have two advantages: detection of variants is greatly simplified to pairwise alignments rather than complicated read-map based inference, which is particularly challenging for indels; and the detection of sequences not present in the reference.

Compared to reference-based approaches, the competitive disadvantage of *de novo* assembly methods is that they disregard the high information content of the reference. Depending on genetic background, greater than 99.5% of anybody's two haplomes is identical to the reference, which therefore constitutes a highly accurate scaffold for personal genomes. It stands to reason that, in principle, an assembly-based method that incorporates information from the reference sequence should combine the advantages of both approaches. We were therefore motivated to develop a new approach to accomplish these three goals via a reference-assisted,

assembly-based approach: high quality diploid personal genome reconstruction; accurate detection of SNPs, indels and SVs; phasing of all types of variants.

Our method, Aquila, makes use of the reference human genome as a reliable scaffold, and performs *de novo* local assembly in small chunks separately for each haplotype, yielding a truly diploid whole genome sequence. It then discovers the most important types of variation on the basis of pairwise alignment to the reference, and infers phasing for all types of assembled variants through previous long-range phasing information. We tested its performance with six libraries of 10x linked-reads data for NA12878 and NA24385 individuals. It offers excellent small indel and SV detection at virtually no compromise for SNP detection, as well as highly accurate phasing of the vast majority of heterozygous variants, at reasonable reagent and computational costs.

Results

Aquila architecture and workflow

Aquila consists of four stages (Figure 1a): Haplotyping and sorting inferred long fragments and their reads into either maternal or paternal bins, locally partitioning reads of each bin, assembling each local partition into sequence contigs, and finally variant calling and phasing.

Each stage has been implemented as a specific module (Figure 1b). In the Haplotyping module, the original long DNA fragments are reconstructed virtually based on their barcodes and on barcode-aware alignment of the reads to the reference sequence. In parallel, SNPs are detected on the basis of the short-read alignments. Fragments are then clustered into either maternal or paternal bins labeled by each pair of heterozygous SNPs, and a probabilistic model (online methods) is applied to exclude clusters caused by sequencing errors. (We here use “maternal” and “paternal” only to distinguish the two parental haplotypes, and do not mean to imply the actual parent of origin.) The clusters are then merged into fewer but larger ones by a greedy recursive algorithm (Supplementary Figure 1), preserving the separation of maternal and paternal bins. The resulting clusters contain all reads from a haplotype block of one parent only and are therefore free of allelic variation, greatly simplifying the later assembly steps.

Before assembly is carried out, haplotype blocks whose lengths exceed a user-defined threshold (default = 200kb) are partitioned into smaller chunks (default = 100kb) by the Contiguity module. This step is necessary because a considerable fraction of the data is in large phase blocks (up to the length of an entire chromosome arm) and handing all these reads to the assembler would partially defeat Aquila's motivation to sidestep the extreme complexity of whole-genome de novo assembly. Thus, the assembly must be broken down into smaller chunks, preferably in places that are highly likely to produce locally correct assemblies, so that contigs from neighboring sub-assemblies can be spliced back together at the partitioning points with high confidence (online methods; Supplementary Figure 2). The locations of the partitioning points are therefore selected such that (1) they are not within repeats in the reference genome and (2) exhibit the expected read coverage and perfect mapping uniqueness in the abovementioned short-read alignments.

The Assembly module (using SPAdes but any appropriate assembler can be used) produces the complete set of "minicontigs" from the partitions, which are then spliced together to produce final contigs (Supplementary Figure 2). While the emphasis in personal genomes is on producing variant calls (see below), we note that traditional short-read-based methods that go directly from read alignment to variant call do not produce this output. The set of final contigs represents a valuable personal resource in the form of an actual genome sequence.

The Variation module then aligns the assembled contigs to the reference. All of the SNPs, small indels, and SVs that Aquila reports are detected on the basis of these pairwise alignments. SNPs are also detected on the basis of these alignments but this assembly-based set is merged with the initial set of SNPs identified on the basis of the barcode-aware read alignment performed by the Haplotyping module. The last step in the Variation module is the phasing of all discovered heterozygous variants. The final output is a VCF with all phased variation that also includes all detected variants for which the individual is homozygous compared to the reference sequence, and the small fraction of heterozygous variants that could not be phased.

Characteristics of Aquila assemblies of six libraries

To explore the performance of Aquila we generated six libraries of 10x linked-read sequencing data from gentle DNA preparations of NA12878 and NA24385 cell lines. The average inferred DNA fragment lengths varied among libraries (Supplementary Table 1), and we numbered them

for each individual according to physical coverage, in ascending order. All of the libraries had approximately 100x Illumina sequencing coverage, except L2 which had 192x. We assembled these libraries with Aquila and compared standard assembly statistics with Supernova 2 assemblies of the same libraries (Table 1, Supplementary Figure 3). Contiguity, as measured by Contig N50, was generally greater than 100kb; NA50, which is the N50 of contigs after breaking at potential misassemblies by comparison to the reference genome, was in most cases less than 5% lower, indicating few misassemblies. We note that using the reference to identify potential misassemblies is a conservative approach given that it is not the same individual as the sequenced one. Contiguity generally increased as a function of physical coverage (C_F; L5 producing the best result at C_F=803x and C_R=0.08), and was greatest with a weighted fragment length of around 150 kb (Supplementary Figure 3).

The fraction of the reference genome covered by the assemblies ranged around 95%, indicating that the vast majority of the non-N and not highly repetitive regions of the genome were covered by all libraries assembled by either Aquila or Supernova2. Aquila consistently produced 98% of this fraction in a diploid state, compared to Supernova's 73-78% (Table 1). This key metric indicates that Aquila produces assemblies that have the potential to support diploid variant detection genome-wide.

Aquila also seamlessly supports combining different libraries, if greater contiguity is desired than is achievable with a single library (Supplementary Figure 4). To illustrate this, we computed the same metrics as for the single libraries for the combination of L5 and L6 from NA24385. Contig N50 increased considerably (by ca. 30-40%) to ca. 178kb, the fraction of reference genome assembled rose to 97%, and the diploid fraction reached 99% of the assembled genome. While assembly statistics represent an important facet for evaluation of Aquila, the ultimate metric for the usefulness of the approach is how well it detects genetic variation. We therefore evaluated the assemblies for their ability to detect SNPs, small indels, and larger structural variants.

Assembly-based detection of SNPs and small indels

We first evaluated assembly-based SNP and small-indel (<50bp) detection by comparing Aquila's calls against the Genome in a Bottle (GiaB) benchmark callsets (Zook et al., 2018). The libraries with the best assembly statistics, L3 (from NA12878) and L5 (from NA24385), achieved 98.8% and 98.7% sensitivity for SNPs and >90% sensitivity for the high-confidence set of GiaB

small indels (Table 2; Supplementary table 2). Genotyping errors for the calls that matched GiaB were 0.14%-0.15% for SNPs and 1.45% to 1.68% for high-confidence small indels. The total numbers of assembly-based SNP calls were 3,967,037 (L3) and 3,878,409 (L5), compared to the total numbers of FreeBayes-based calls performed on the barcode-aware read alignments of 3,949,721 (L3) and 3,961,684 (L5). Numbers of heterozygotes or homozygotes are also comparable between the two approaches (Supplementary Table 3). Due to the lack of true gold standard genome sequences of NA12878 and NA24385, we are unable to determine the specificity of calls or the false-positive rates, but we note that Aquila produces numbers of assembly-based SNP calls for these two individuals that are consistent with those previously produced from standard short-fragment Illumina libraries (Wu et al., 2017; Li et al., 2018; Supernat et al., 2018).

Compared to the GiaB small indel callset, Aquila produces considerably more calls (e.g., 870,990 in L3 vs GiaB's 531,228; Supplementary Table 4). This difference is due to the current incompleteness of the GiaB callset especially among longer small indels, as well as a false-positive rate of the Aquila calls that we cannot rigorously estimate. The size distribution of Aquila's small indels matches the size distribution of the GiaB calls 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 (Supplementary Figure 5). At lengths above 30bp, where GiaB has very few calls, the Aquila calls continue to exhibit this pattern. The correlations between the Aquila and GiaB distributions of the 1-49 bp indel calls are R^2=0.997 and 0.998 of the raw counts, and 0.930 and 0.951 for natural log of counts, for insertions and deletions, respectively.

Assembly-based detection of structural variants 50bp and greater

Aquila calls ca. 10,000 deletions and ca. 6,000 insertions 50bp and greater in each high quality library (L2, 3, 5, and 6; Supplementary Table 5). The size distributions follow the expected exponential distribution with a peak at ca. 330 bp, which is caused by full-length or nearly full-length Alu elements (Figure 2a). The number of calls is comparable but consistently lower than recent estimates from a comprehensive study that focused exclusively on SVs in a cohort of long-read sequenced individuals, including two haploid samples, in which purpose-driven approaches were applied to achieve high sensitivity of detection of shared SVs (Audano et al., 2019; Chaisson et al., 2019). We do not expect to reach the same level of detection in a single

personal genome without the benefit of leveraging several individuals to inform discovery, but we are able to apply several metrics to characterize Aquila's SV calls (Figure 2b-g).

We used a combination of three strategies to validate the SV calls in both individuals: First, for both individuals, PacBio data exist that we applied with svviz2 (Spies et al., 2015) to test Aquila's calls with another data type; second, the two individuals are expected to share a large fraction of SVs, so we performed simple comparisons of the call sets; third, we aligned the SVs and flanking sequence to two high quality ape genomes (Chimp and Orang). Given the complexity of SV calling we expect to have both false negative and false positive calls, which is underscored by the fact that different libraries from the same individual produce strongly overlapping but not fully identical call sets. The calls that are shared between libraries validate (by at least one of the abovementioned three metrics) at 92% for NA12878 and 82% for NA24385 (Figure 2b), whereas calls unique to each library validate at lower rates (40% to 52%). This shows that we do not have perfect sensitivity (because there are many validated calls unique to a single library) and, conversely, that the nonvalidated set of calls unique to a library likely contains false positives.

The number of SVs validated by svviz2/PacBio is higher than those validated by the other two approaches (Figure 2c), since the PacBio data is from the same individual as the respective assembly. However, there are many SVs that are not validated by PacBio but that are also called in the other individual or are validated by comparison to ape sequences. This effect is greater for deletion calls than for insertion calls (Figure 2c), which suggests that insertions may be called at higher specificity and lower sensitivity than deletions. This interpretation is consistent with a predicted shortcoming of the current implementation of Aquila, which because of its reliance on the reference sequence to identify reads for assembly has decreasing power to assemble insertions as their size increases.

We also assessed the consistency of the breakpoints of the calls that are shared between the two individuals (Figure 2d,e). We binned the SVs based on the size differences between the calls in NA12878 and NA24385, as a function of validation by the other two approaches (Figure 2d) or as a function of sequence type (Figure 2e). Overall, the vast majority of calls have precisely the same breakpoints in both individuals. Deletions or validated calls have better precision than insertions or nonvalidated calls, and SVs in repeats have worse precision than nonrepetitive sequences.

Inference of derived alleles and comparison of genotype frequencies

SV calls are labeled with respect to the reference sequence as 'insertion' or 'deletion', but the molecular mechanism that generated the SV may be the opposite of the call because the reference sequence is a random sample of ancestral and derived alleles from the population. Where the reference carries a derived allele, the actual molecular mechanism that generated the SV is the opposite of the call because the assembly carries the ancestral allele. For those SVs with alignments to ape genomes, the SV allele matching the ape sequence is highly likely to represent the ancestral state, and the other allele is derived, allowing inference of the actual molecular mechanism that generated it (Figure 2f). Classifying the SVs accordingly ("actual" insertion or deletion) causes a striking shift in the size distributions of insertions and deletions in both individuals (Figure 2a,g), in which the vast majority of SVs that overlap Alu repeat sequences are now revealed to be actual insertions (Figure 2g). This provides empirical evidence that the classification into ancestral and derived alleles is largely correct, as Alus are known to insert as full length sequences, whereas partial Alus are degenerate copies that arise later in evolution from deletions whose breakpoints can be anywhere in the element.

We also asked whether segregation of the SV alleles behaves as expected. For each locus, we have two chromosomes each from the sequenced individuals and one from the reference genome, for a total of 5, and therefore 18 possible combinations of ancestral and derived alleles among them (Figure 3a). We cannot observe loci in which all chromosomes are identical, leaving 16 patterns that contain one to four derived alleles. Population genetic principles state that derived alleles have systematically lower allele frequencies than ancestral ones. Indeed, for both insertions and deletions, the SV loci that have one derived allele (and four ancestral ones) are much more common than those that have 2 or more (Figure 3b). In 12 of the segregation patterns, genotypes of NA12878 and NA24385 differ. These can be arranged in 6 pairs where the individuals' genotypes are equivalent, for example, one heterozygote and one ancestral homozygote, with the reference carrying a derived allele (rightmost green box in Figure 3a). Distinguishing insertions and deletions, this gives 12 classes (bottom of Figure 3a), each of which has two SV counts that should be very similar to each other assuming that the two individuals do not come from very different populations. Indeed, the correlation between these equivalence classes is very high (R^2=0.92, Figure 3c,d).

Genome-wide distribution and phasing of all variation

To interrelate the different types of variation detected by Aquila and to ask whether there were any obvious biases we divided the genome into bins of 250kb and quantified tandem repeat content, contig density, and variation content by genotype from the assembly of library 3 (Figure 4a). Contig density, which over the vast majority of the genome is exactly 2 because of the diploid nature of the assemblies, does not correlate with any variation and only weakly with repeat content. Repeat content correlates weakly with the number of SVs. As expected, numbers of SNPs and very small indels correlate strongly when the genotype is the same (heterozygous or homozygous), and weakly with larger small indels, which in turn correlate weakly with SVs. Overall, the correlation patterns do not reveal any large-scale biases in variation discovery. We also note that the fraction of variants that are heterozygous varies over a narrow range across all types and sizes of detected variation (Supplementary Figure 6), again revealing no obvious biases.

The last step in Aquila's pipeline is a final integrative phasing of all of the discovered heterozygous variation on the basis of the phase blocks obtained with heterozygous SNPs in the Haplotyping module (Supplementary Table 6). Depending on the library, between ca 1.7 and 2 million SNPs were initially phased (Supplementary Table 7). Because the parental genotypes are known for the two individuals we could quantify the phasing error, which is dominated by switch errors that involve a single, presumably incorrectly genotyped, SNP. Long switch errors are quite rare, comparing favorably with previous work using 10x data (Shajii et al., 2018; Marks et al., 2019). Because SNPs are the densest type of polymorphism in human genomes, phasing other variants on the basis of these is feasible. Instead of probabilistic imputation, however, Aquila performs straightforward inference by matching the assembly-based SNP calls with those of the Haplotyping module and then simply inferring the correct phase (online methods; Supplementary Figure 7). In total, Aquila could add ca. 1 million heterozygous variants, including 0.5 – 0.8 million previously unphased SNPs, ca. 0.5 million small indels, and ca. 10,000 heterozygous SVs in the best four libraries (Supplementary Table 7).

Discussion

We introduce a new method, Aquila, which performs *de novo* assembly in small chunks for both haplotypes, producing a truly diploid personal genome sequence from a single data type,

Illumina/10X linked reads. We test Aquila's performance on several libraries from two standard individuals and show that it produces high quality diploid assemblies with accurate phasing. The assemblies enable comprehensive discovery of SNPs, small indels, and SVs, on the basis of pairwise alignments to the reference genome.

Aquila is the first approach that effectively leverages the strengths of 10X/Illumina sequencing to enable comprehensive variation discovery and phasing in personal genomes at a reasonable cost. Compared to standard Illumina sequencing, the increased cost of the 10X sample prep and the generation of deeper sequence data is justified by (1) the much greater power to detect insertions and deletions, and (2) the genome-wide phasing of all heterozygous variants. Compared to ONT and standard PacBio, the superior base-pair level accuracy of Illumina sequencing ensures more accurate SNP and small indel detection as well as SV breakpoint determination; and while PacBio CCS appears to have the potential to enable highly accurate genome sequencing it remains to be seen whether it performs so much better than 10X/Illumina/Aquila to justify the additional cost, particularly for cohort studies. Compared to ensemble approaches that computationally integrate multiple data types with complementary strengths (for example, combining Illumina, long-read, and BioNano data), 10X/Illumina/Aquila is far less complex to manage in the laboratory and offers a simpler computational approach.

In the ecosystem of solutions for human whole genome sequencing, 10X/Illumina/Aquila therefore fills what we believe is the most important niche: a diploid and phased personal genome for accurate and comprehensive discovery of SNPs, small indels, and SVs in all but the most complicated regions of the human genome. It represents the first generation of approaches that drive toward laboratory and computational efficiency and simplicity by using a single data type and leveraging the considerable amount of information present in the human reference sequence. Until de novo assembly on the basis of highly accurate very-long-read data is shown to be cost-effective, reference-assisted approaches that partition the genome into smaller assembly problems are likely to prevail. Future improvements of Aquila will center on better detection of long insertions and contig breakpoint assembly. Although developed for 10X/Illumina data, Aquila's architecture may be used in the future for computational approaches that use data from other sources. In its current form, it is already applicable to any study that requires better indel detection than what is achievable with standard Illumina sequencing.

Software availability: Aquila can be found at https://github.com/maiziex/Aquila. Version 1.1 was used to generate the results in this paper.

Data availability: Assemblies and VCFs can be found at http://mendel.stanford.edu/supplementarydata/zhou_aquila_2019/. We will be submitting raw sequence data and assemblies to NCBI's SRA and Assembly databases.

Acknowledgements

This research was supported by the Joint Initiative for Metrology in Biology (JIMB; National Institute of Standards and Technology). We would like to thank Noah Spies, Justin Zook, and Marc Salit for informative discussions.

NA12878	Raw Coverage C (X)	C_F (X)	C_R (X)	Contig N50 (bp)	Contig NA50 (bp)	Diploid Fraction (%)	Genome Fraction (%)
L1_Aquila	103	123	0.41	34,759	31,645	98.1	95.45
L1_Supernova				58,408	57,254	73.3	93.69
L2_Aquila	192	334	0.27	94,699	91,662	96.7	94.43
L2_Supernova				144,298	136,253	58.9	94.81
L3_Aquila	106	958	0.07	120,963	116,438	98.7	96.17
L3_Supernova				114,900	111,388	77.2	94.62
NA24385							
L4_Aquila	100	208	0.25	96,558	93,132	98.2	96.98
L4_Supernova				99,516	96,510	79.2	95.01
L5_Aquila	100	803	0.08	135,115	130,176	98.8	96.77
L5_Supernova				129,195	125,339	78.1	94.86
L6_Aquila	100	1504	0.05	125,064	120,490	98.7	96.62
L6_Supernova				101,236	98,364	73.4	94.69
L5+L6_Aquila	200	2307	0.07	177,974	172,358	99.1	97.06

Table 1: Assembly metrics for the six libraries we built from NA12878 and NA24385. Each row represents statistics for both Aquila and Supernova2, for comparison. C_F, physical coverage; C_R, read coverage; C, Raw Coverage >= C_F x C_R. 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. L5+L6 describes performance for a simple combination of the data from libraries 5 and 6.

	Calls		Genotype error of correct calls			
	Number of assembly-based calls in GiaB benchmark set	**Sensitivity (%)**	**Het (G) => Hom**	**Hom (G) => Het**	**Total G => Aq**	**Error (%)**
12878 L3						
SNPs	3,047,725 / 3,084,732 =	98.800	3,524	1,187	4,711	0.1546
Indels	452,593 / 497,256 =	91.018	3,795	3,827	7,622	1.684
24385 L5						
SNPs	3,036,939 / 3,076,552 =	98.712	3,082	1,143	4,225	0.1401
Indels	428,050 / 473,761 =	90.351	3,298	2,910	6,208	1.450

Table 2: Sensitivity of assembly-based detection of SNP and small Indels (< 50bp) for library 3 of NA12878 and library 5 of NA24385. Last four columns are the number and rate of incorrect genotypes for the set of variants in GiaB that Aquila also calls.

Figure Legends

Figure 1. Aquila architecture. Lowercase letters in circles denote existing programs we integrated (a = LongRanger Align, b = FreeBayes, c = SPAdes). (a) Overall architecture. (b) Detailed workflow. Green boxes are data, arrows indicate input and output of a pipeline component. Input data are a high quality reference genome and 10X-based short reads, each with a barcode (not shown). The Haplotyping module produces phased virtual long fragments (by alignment of reads to the reference, SNP detection, and haplotyping) that become part of each read's record. The Contiguity module produces specific single-base coordinates ('partitioning points') in the genome and in the data, where haplotype blocks and reads are cut at a specific single location, and where subsequently assembled minicontigs are rejoined in the end. The Local Assembly module executes assembly of the reads of a specific region, separately for each parental copy. The variation module then discovers, integrates, and infers the phase of all variation.

Figure 2. Characteristics and validation statistics of SV calls by Aquila. (a) Size frequency distributions of insertion calls and deletion calls in both individuals (L3 for NA12878 and L5 for NA24385). Black areas represent indels of which at least 80% are a close match to Alu-element consensus sequences. (b) Call validation rates by three validation strategies of two libraries (L2 and L3; L5 and L6) per individual; SVs called in both libraries are in the overlap, flanked by SVs unique to each library. (c) Overlap analysis and comparison of three validation strategies, by call and individual; numbers inside the Venn diagrams are counts of SVs. SVs are validated by: PacBio data from the same individual, (PacBio); the other individual (Both Individuals); in the chimp or orang genome (Apes). Overlaps represent two or more of these criteria fulfilled. (d,e) Comparative precision of SVs present in both individuals, as a function of validation by three validation strategies (d) or sequence class (e). Bar graphs depict counts of SVs that have precisely the same breakpoint coordinates in both individuals (0 bp), that differ by less than 10 (1-9 bp), or that differ by 10 or more (>=10 bp). "Repeats" class includes simple sequence and tandem repeats but not mobile elements; "Other" class includes all SVs that do not overlap more than 80% with Alus and are not part of the Repeats class. (f) Inference of actual molecular mechanism that produced the SV by expanding the alignment between the reference sequence (Ref) and the Individual (Ind) to include chimp or orang sequences; the sequence that matches the ape is the ancestral allele. "Actual insertion" and "Actual deletion" refer to the molecular mechanism that produced the derived allele. Approximately 45% of deletion and 24% of

insertion calls are thus 'inverted' (blue arrows). (g) Size frequency distributions of actual insertions and actual deletions in both individuals. Black areas represent indels of which at least 80% are a close match to the Alu-element consensus sequence. The peak at around 330 base pairs captures nearly all Alu SVs.

Figure 3. Frequencies and genetic configurations of detected actual insertions and deletions as determined by alignment to ape genomes. (a) Each column depicts a possible configuration of ancestral and derived alleles in the five haplomes; all possibilities are shown but we cannot detect variation shown in the first and last columns as all sequences are identical in those cases. Each pattern is distinct but, due to the nature of sampling, represents between 1 and 4 possible arrangements. The number of derived alleles in each pattern is also indicated. Shorthand for symbols indicating the molecular type and who carries the derived alleles (MD) are: DT, deletion carried by a target; DR, deletion carried by the reference; IT, insertion carried by a target, IR, insertion carried by the reference. 0,1: single heterozygote; 0,2: single homozygote; 1,2: one target is a heterozygote, the other a homozygote. (b,c) MD numbers in the individuals correlated to each other. Symbols represent classes as described in (a). For example, the green diamond (IR 0,1) indicates SVs defined by the fact that the derived allele is an insertion ("I") in the reference sequence ("R"), designated as "IR". There are 475 SVs in which NA12878 is homozygous for the reference allele and NA24385 is heterozygous for the derived allele, and 428 SVs for which NA12878 is heterozygous for the derived allele and NA24385 is homozygous for the reference allele. All other data points are analogous, always comparing NA12878 calls to their independent and equivalent counterparts.

Figure 4. Local distribution of all types of variation detected, aggregated across 250kb intervals. (a) Top, ideogram of Chromosome 1 showing metaphase banding pattern. Tracks below are assembly and variation features where white represents no data. Repeats, Z-score of TRF-detected repeat density. Contigs, average number of contigs per base pair. SNPs, Z-score of SNP number. Small indels, Z-score of the number of indels 1-50bp. SVs, number of SVs >=50bp in each 250kb bin. Phase blocks, each phase block is a grey rectangle, with alternating light and dark indicating neighboring phase blocks. (b) Genome-wide correlation (R^2) among all pairs of variation types by genotype, contig density, and repeat density.

References

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, Warren WC, Magrini V, McGrath SD, Li YI, Wilson RK, Eichler EE (2019) Characterizing the Major Structural Variant Alleles of the Human Genome. Cell 176:663-675.e19 Available at: http://www.ncbi.nlm.nih.gov/pubmed/30661756 [Accessed May 26, 2019].

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 19:455–477 Available at: http://www.liebertpub.com/doi/10.1089/cmb.2012.0021 [Accessed September 24, 2018].

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580 Available at: http://www.ncbi.nlm.nih.gov/pubmed/9862982 [Accessed February 8, 2019].

Chaisson MJP et al. (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun 10:1784 Available at: http://www.ncbi.nlm.nih.gov/pubmed/30992455 [Accessed May 26, 2019].

Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC (2016) Phased diploid genome assembly with single-molecule real-time sequencing. Articles 1050 Available at: https://www.nature.com/articles/nmeth.4035.pdf [Accessed September 26, 2018].

Coster W De, Roeck A De, Pooter T De, D'Hert S, Rijk P De, Strazisar M, Sleegers K, Broeckhoven C Van (2018) Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. bioRxiv:434118 Available at: https://www.biorxiv.org/content/10.1101/434118v1 [Accessed May 20, 2019].

Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, Pregno G, Giachino D, Mandrile G, Espejo Valle-Inclan J, Korzelius J, de Bruijn E, Cuppen E, Talkowski ME, Marschall T, de Ridder J, Kloosterman WP (2017) Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Commun 8:1326 Available at: http://www.nature.com/articles/s41467-017-01343-4 [Accessed May 9, 2019].

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797 Available at: http://www.ncbi.nlm.nih.gov/pubmed/15034147 [Accessed February 8, 2019].

Fan X, Chaisson M, Nakhleh L, Chen K (2017) HySA: a Hybrid Structural variant Assembly approach using next-generation and single-molecule sequencing technologies. Genome Res 27:793–800 Available at: http://www.ncbi.nlm.nih.gov/pubmed/28104618 [Accessed May 9, 2019].

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075 Available at: http://www.ncbi.nlm.nih.gov/pubmed/23422339 [Accessed February 8, 2019].

Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, Peluso P, Boitano M, Chin C-S, Korlach J, Wilson RK, Eichler EE (2017) Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res 27:677–685 Available at: http://www.ncbi.nlm.nih.gov/pubmed/27895111 [Accessed May 9, 2019].

Jiang T, Huang M, Jiang T, Gu Y, Wang Y, Wu Y, Ma H, Jin G, Dai J, Hu Z (2018) Genome-wide compound heterozygosity analysis highlighted 4 novel susceptibility loci for congenital heart disease in Chinese population. Clin Genet 94:296–302 Available at: http://www.ncbi.nlm.nih.gov/pubmed/29774522 [Accessed May 30, 2019].

Karimzadeh M, Ernst C, Kundaje A, Hoffman MM (2018) Umap and Bismap: quantifying genome and methylome mappability. Nucleic Acids Res 46 Available at: https://bismap.hoffmanlab.org [Accessed June 4, 2019].

Larsen PA, Heilman AM, Yoder AD (2014) The utility of PacBio circular consensus sequencing for characterizing complex gene families in non-model organisms. BMC Genomics 15:720 Available at: http://www.ncbi.nlm.nih.gov/pubmed/25159659 [Accessed May 9, 2019].

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences Birol I, ed. Bioinformatics 34:3094–3100 Available at: http://www.ncbi.nlm.nih.gov/pubmed/29750242 [Accessed February 8, 2019].

Li Z, Wang Y, Wang F (2018) A study on fast calling variants from next-generation sequencing data using decision tree. BMC Bioinformatics 19:145 Available at: http://www.ncbi.nlm.nih.gov/pubmed/29673316 [Accessed May 30, 2019].

Lupski JR et al. (2010) Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy. N Engl J Med 362:1181–1191 Available at: http://www.nejm.org/doi/abs/10.1056/NEJMoa0908094 [Accessed May 26, 2019].

Marks P et al. (2019) Resolving the full spectrum of human genome variation using Linked-Reads. Genome Res 29:635–645 Available at: http://www.ncbi.nlm.nih.gov/pubmed/30894395 [Accessed May 26, 2019].

Nattestad M, Schatz MC (2016) Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics 32:3021–3023 Available at: http://www.ncbi.nlm.nih.gov/pubmed/27318204 [Accessed May 20, 2019].

Nattestad M, Schatz MC (n.d.) Assemblytics: a web analytics tool for the detection of assembly-based variants. Bioinformatics, YYYY:0–0 Available at: https://www.biorxiv.org/content/biorxiv/early/2016/03/20/044925.full.pdf [Accessed February 8, 2019].

Redon R et al. (2006) Global variation in copy number in the human genome. Nature 444:444–454 Available at: http://www.nature.com/articles/nature05329 [Accessed May 9, 2019].

Rhoads A, Au KF (2015) PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics 13:278–289 Available at: http://www.ncbi.nlm.nih.gov/pubmed/26542840 [Accessed May 9, 2019].

Shajii A, Numanagić I, Berger B (2018) Latent Variable Model for Aligning Barcoded Short-Reads Improves Downstream Analyses. Res Comput Mol Biol . Annu Int Conf RECOMB . proceedings RECOMB (Conference 2005-) 10812:280 Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5989713/ [Accessed October 4, 2018].

Spies N, Weng Z, Bishara A, McDaniel J, Catoe D, Zook JM, Salit M, West RB, Batzoglou S, Sidow A (2017) Genome-wide reconstruction of complex structural variants using read clouds. Nat Methods 14:915–920 Available at: http://www.ncbi.nlm.nih.gov/pubmed/28714986 [Accessed May 26, 2019].

Spies N, Zook JM, Salit M, Sidow A (2015) Svviz: A read viewer for validating structural variants. Bioinformatics 31:3994–3996 Available at: http://www.ncbi.nlm.nih.gov/pubmed/26286809 [Accessed May 21, 2019].

Supernat A, Vidarsson OV, Steen VM, Stokowy T (2018) Comparison of three variant callers for human whole genome sequencing. Sci Rep 8:17851 Available at: http://www.ncbi.nlm.nih.gov/pubmed/30552369 [Accessed May 30, 2019].

Travers KJ, Chin C-S, Rank DR, Eid JS, Turner SW (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res 38:e159 Available at: http://www.ncbi.nlm.nih.gov/pubmed/20571086 [Accessed May 9, 2019].

Wala JA, Bandopadhayay P, Greenwald N, O'rourke R, Sharpe T, Stewart C, Schumacher S, Li Y, Weischenfeldt J, Yao X, Nusbaum C, Campbell P, Getz G, Meyerson M, Zhang C-Z, Imielinski M, Beroukhim R (2018) SvABA: genome-wide detection of structural variants and indels by local assembly. Available at: http://www.genome.org/cgi/doi/10.1101/gr.221028.117. [Accessed February 8, 2019].

Wang O et al. (2019) 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 Res 29:798–808 Available at: http://www.ncbi.nlm.nih.gov/pubmed/30940689 [Accessed May 9, 2019].

Weischenfeldt J, Symmons O, Spitz F, Korbel JO (2013) Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet 14:125–138 Available at: http://www.ncbi.nlm.nih.gov/pubmed/23329113 [Accessed May 9, 2019].

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB (2017) Direct determination of diploid genome sequences. Genome Res 27:757–767 Available at: http://www.ncbi.nlm.nih.gov/pubmed/28381613 [Accessed February 7, 2019].

Wu L, Yavas G, Hong H, Tong W, Xiao W (2017) Direct comparison of performance of single nucleotide variant calling in human genome with alignment-based and assembly-based approaches. Sci Rep 7:10963 Available at: http://www.ncbi.nlm.nih.gov/pubmed/28887485 [Accessed May 30, 2019].

Zhang L, Zhou X, Weng Z, Sidow A (n.d.) De novo diploid genome assembly for genome-wide structural variant detection. Available at: http://dx.doi.org/10.1101/552430 [Accessed February 17, 2019].

Zheng GXY et al. (2016a) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat Biotechnol 34:303–311 Available at: http://www.ncbi.nlm.nih.gov/pubmed/26829319 [Accessed February 7, 2019].

Zheng GXY et al. (2016b) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat Biotechnol 34:303–311 Available at: http://www.ncbi.nlm.nih.gov/pubmed/26829319 [Accessed May 9, 2019].

Zook JM, McDaniel J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean CY, Vega FMD La, Xiao C, Sherry S, Salit M (2018) Reproducible integration of multiple sequencing datasets to form high-confidence SNP, indel, and reference calls for five human genome reference materials. bioRxiv:281006 Available at: https://www.biorxiv.org/content/10.1101/281006v2.full [Accessed May 30, 2019].

ONLINE METHODS

Aquila is organized in four conceptual modules that correspond to the following python steps:

Aquila_step1.py: **Haplotyping** module + contiguity module (first part: **partitioning**)

Aquila_step2.py: Local **assembly** module + contiguity module (second part: **concatenation**)

Variation Module: Aquila_assembly_based_variants_call.py and Aquila_phasing_all_variants.py

Pruning unreliable variants in the Haplotyping module

Correct haplotyping requires filtering out incorrectly genotyped variants and false positives due to sequencing error. We performed an empirical analysis for 10x sequencing data to investigate the alternate allele frequency ($R_{alt/ref}$) and coverage per variant (d_{var}) that could be used as metrics to find erroneous calls. Allele frequency ($R_{alt/ref} >= 0.25$) was used for a cutoff, and a 2-tailed percentile cutoff was used for coverage per variant ($10\%*avg_cov <= d_{var} <= 90\%*avg_cov$, *avg_cov*: average read coverage per variant). The haplotyping algorithm was further improved by sacrificing a small amount of low-confidence heterozygous variants. SNP quality (*13* by default) is the final free parameter used to prune variants.

Original molecule (long-fragment) haplotyping and phasing accuracy

All DNA fragments are first reconstructed by aligning short reads to the human genome reference (Hg38) by a barcode-aware alignment strategy ('*Longranger align',* https://support.10xgenomics.com/genome-exome/software/pipelines/latest/installation). Aquila then sorts all reads by barcodes and positions, and collects the reads with the same barcode to reconstruct each fragment. There is a boundary threshold to differentiate two molecules with the same barcode when the distance between two successive reads with the same barcode is larger than 50kb (50kb by default, free parameter). After reconstructing all fragments, Aquila assigns the alleles of heterozygous SNPs to each fragment by scanning the reads belonging to each fragment and comparing to a VCF file generated by FreeBayes. At a heterozygous locus "*0*" is the reference allele and "*1*" is the alternate allele.

For each pair of heterozygous variants, if the even parity was correct where one haplotype supported "*00*", and the other haplotype supported "*11*", the odd parity must then have been caused by a sequence error with some fragments supporting "*01*", and other fragments supporting "*10*", and vice versa. In rare scenarios, the fragment could have two sequencing

errors when the even/odd parity was correct, but the fragment supported the complementary haplotype (e.g., the haplotype is ""*00*"", and the fragment supports "*11*"). For each fragment with at least two heterozygous SNPs, Aquila records all neighboring pairs of heterozygous variants. It then applies a Bayesian model (see below) to evaluate if even or odd parity is correct, and the clusters with the parity caused by sequence error are excluded from further steps. Importantly, the excluded clusters are due to the variants caused by sequencing errors, not the molecules themselves, which means if these molecules still contain other pairs of heterozygous variants with consistent haplotype with the correct parity, they are still used for haplotyping.

Aquila then performs a recursive clustering algorithm (Supplemental Figure 1) in two haplotypes to aggregate bigger clusters/phase blocks. Two clusters are merged if the number of molecules in both of them supporting the same haplotype exceeds a threshold. This threshold is set to 3 by default, which corresponds 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 p_1 and p_2, respectively. Aquila sorts all pairs of clusters by the positions of reads of all molecules in each cluster. When two locally successive clusters are merged into one single cluster, the corresponding clusters of the other haplotype are merged too. The resulting pairs of clusters are sorted again for the next iteration. The sorting algorithm complexity is $\sim O(N_{var} log N_{var})$, where N_{var} is the total number of heterozygous variants. Aquila performs clustering recursively until no more clusters can be merged based on the supporting threshold.

The result of this step are pairs of clusters where each pair corresponds to one diploid phase block. For each phase block, Aquila then performs haplotype construction and extension when each heterozygous variant is supported by all the molecules that cover it. When there are multiple molecules supporting inconsistent genotypes for a variant, that variant is excluded from further steps. To further extend the phase blocks, Aquila similarly performs recursive clustering when two phase blocks have a number of overlapping variants greater than a certain 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.

In the final step, all the molecules/fragments with at least two heterozygous variants are assigned to a phase block based on the variants in the final phase block, and a maximum likelihood estimation is applied. Given a haplotype *H* and a molecule *M*, we apply the theta

function $\theta(H_i, M_i) = 1$ if $H_i = M_i$ and *0* otherwise. Given p_i, the probability the allele call at variant *i* in molecule *M* is correct, the likelihood of observing molecule *M* is:

$$p(M|p, H) = \Pi\ \theta(H_i, M_i)p_i + (1 - \theta(H_i, M_i))(1 - p_i)$$

Similarly, given the complementary haplotype H_c and molecule *M*, a theta function gets the value $\theta(H_{ci}, M_i) = 1$ if $H_{ci} = M_i$ and *0* otherwise. The likelihood of observing molecule *M* is: $p(M|p, H_c) = \Pi\ \theta(H_{ci}, M_i)(1 - p_i) + (1 - \theta(H_{ci}, M_i))p_i$. To assign the final phase block to each molecule *M*, Aquila needs to find the haplotype *j* in the final phase blocks that meets $argmax_j \left(p\left(M\middle|p, H^j\right) - p\left(M\middle|p, H_c^j\right)\right)$.

Haplotyping/phasing accuracy is defined by switch errors (Supplemental Table 6). In Aquila's haplotyping algorithm, the goal is to improve phasing accuracy by pruning low-confidence variants, instead of aggressively extending phase blocks, which sacrifices accuracy. SNPs that are not phased here still have a chance of getting phased if they are 'rediscovered' during assembly-based variant finding.

Probability model to determine correct joint key

At each position, the genome has two complementary "keys" matching the true two haplotypes: *00*, *11* (even parity) or *01*, *10* (odd parity). Each variant matches the true variant with probability p_1 and p_2, respectively. The probability that a sequence key will have the correct parity is $p_c = p_1p_2+(1-p_1)(1-p_2)$, since both variants could match the true variants or both variants are called wrong. Let *N* be the number of sequences observed that have both variants, and *k* be number of sequences with key of even parity, and *(N − k)* be the number of sequences with key of odd parity. Using Bayes' theorem to test *P(B | A)*:

Where *A*: *k* out of *N* molecules have keys with even parity, *B*: true key is even parity.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Aquila will accept even parity is correct if *P(B | A)* exceeds a significance level (e.g. > *0.99*), where:

P(B | A) is the probability of the true key being even parity given *k* out of *N* molecules have keys with even parity.

P(A | B) is the probability of *k* out of *N* molecules having keys with even parity given the true key is even parity, which is $P(A|B) = \binom{N}{k} \cdot p_c^k (1 - p_c)^{N-k}$.

P(A) is the probability of k out of N molecules have keys with even parity, which is $P(A) = \binom{N}{k} / \left(\binom{N}{1} + \binom{N}{2} + \cdots + \binom{N}{k} + \cdots + \binom{N}{N} \right) = \binom{N}{k} / 2^N$.

P(B) is the probability of even parity being correct, *P (B) = 1.*

Phase block splitting

For code optimization, the original molecule / long fragments file is sorted by the start coordinates of all molecules, and if the start coordinate of the current molecule ($mole_i$, i = current position) is larger than the end coordinate of the previous molecule ($mole_{i-1}$), the two blocks are split at position i. The algorithm complexity is ~ $O(N_b L_b \log L_b)$ (N_b: number of blocks, L_b: the block size).

High-confidence partitioning point profile generation

Large phase blocks (free parameter: *block_threshold = 200kb by default*) are cut into multiple small chunks of a specific length (free parameter: *block_len_use = 100kb by default*) based on a high-confidence partitioning point profile. This is done to make assembly faster and more tractable, by avoiding too many reads being given to the assembler. This profile is generated based on three criteria: 1. Expected reads coverage (C), 2. Expected physical coverage (C_F), 3. *100*-mer uniqueness. Read depth and fragment/physical coverage for each position is calculated after reconstructing all the fragments. The 100-mer uniqueness file (Karimzadeh et al., 2018) for hg38 processed and included in Aquila, was downloaded from http://hgdownload.soe.ucsc.edu/gbdb/hg38/hoffmanMappability/k100.Unique.Mappability.bb Each locus/position is defined to be a high-confidence partitioning point if $C_{(partitioning\ point)} > C_{(average)} * 0.8$, $C_{F\ (partitioning\ point)} > C_{F\ (average)} * 0.8$, and *locus* $\in$ *100-mer* uniqueness. Aquila uses these high-confidence partitioning points in the profile as reference points to partition reads before assembly (see next section) and later to reconnect the resulting mini-contigs into contigs (Supplementary Figure 2).

Algorithm for local assembly within small diploid chunks and stitching contigs

For each phase block (Supplemental Figure 2), Aquila records its original long molecules and their corresponding short reads. Large phase blocks are cut into small chunks (see previous step) to perform local assembly with SPAdes (Bankevich et al., 2012) within each phase block for both maternal and paternal haplotypes separately. Those resulting minicontigs from neighboring small chunks that are bounded by the same partitioning point are concatenated.

99% of partitioning points met this criterion. For each concatenating iteration, the previous concatenated contig is used for the next iteration of concatenation. At the end of this step, Aquila has generated contigs for both maternal and paternal haplotypes in each original phase block. The algorithm complexity is ~ $N_{chunks}O(T_{onechunk})$, where N_{chunks} = number of small chunks and $T_{onechunk}$ = time for finishing assembly of one small chunk.

Combining multiple libraries of linked-reads into a single assembly

Each library of 10x linked-read sequencing data uses the same set of barcodes, which makes the combination of multiple libraries at the stage of raw fastq reads difficult. To combine multiple libraries, Aquila reconstructs the original fragments with their corresponding reads of each library separately, after which barcodes are not required any more. All the DNA fragments from multiple libraries, and their corresponding reads, are then combined to perform haplotyping and all subsequent steps. The molecule haplotyping algorithm is efficient and makes haplotyping fragments for multiple libraries fast. The algorithm complexity is linear with the number of libraries, ~ $N_{libs}O(T_{onelib})$, where N_{libs} = number of libraries and T_{onelib} = time for finishing assembly of one library.

Contig quality assessment

We used QUAST (Gurevich et al., 2013) to generate various assembly metrics such as N50 and NA50. *--extensive-mis-size* 1000 was applied as the lower threshold of the relocation size.

Aquila computation cost

Aquila's four modules vary in their memory requirements and are flexible with respect to computing architecture, allowing jobs to be run in parallel on a cluster or serially in a large-memory machine. Generally the run time is considerably longer on the large memory machine. On a cluster, jobs are parallelized by chromosome or a combination of chromosomes that minimizes the number of nodes required. Modules 1 to 3 use 13 + 10, and module 4 use 9 + 6 hours of wall clock time respectively on a current-generation standard compute cluster (12 nodes with 128Gb RAM each, 2 CPUs each, 10-cores per CPU, i.e. 240 cores).

Aquila output

Aquila outputs an overall contig file “Aquila_contig.fasta”, and separately for each chromosome “Aquila_Contig_chr*.fasta”. After performing assembly-based variant calling in the Variation

Module, one contig file for each haplotype is generated: “Aquila_Contig_chr*_hp1.fasta” and “Aquila_Contig_chr*_hp2.fasta”. For each contig, the header (for example “*>36_PS39049620:39149620_hp1*”) follows the format contig number (“*36*”), phase block start coordinate (“*PS39049620*”), phase block end coordinate (“:*39149620*”), and haplotype number (“*hp1*”). Within the same phase block, the haplotype number “*hp1*” and “*hp2*” are arbitrary for maternal and paternal haplotypes. For some contigs from large phase blocks, the headers are much longer and complex, for an instance, *“>56432_PS176969599:181582362_hp1_ merge177969599:178064599_hp1-177869599:177969599_hp1*”. “*56*” denotes contig number, “*176969599*” denotes the start coordinate of the final big phase block, “*181582362*” denotes the end coordinate of the final big phase block, and “hp1” denotes the haplotype “*1*”. “*177969599:178064599_hp1*” and “*177869599:177969599_hp1*” mean that this contig is concatenated from minicontigs in small chunk (start coordinate: *177969599*, end coordinate: *178064599*, and haplotype: *1*) and small chunk (start coordinate: *177869599*, end coordinate: *177969599*, and haplotype: *1*).

Variation detection method for assembled contigs in the Variation Module

To generate single nucleotide polymorphisms (SNPs), small Indel, and structural variant (SV) calls from the *de novo* assemblies, Aquila uses the contig file of the maternal and paternal haplotype of each phase block. Minimap2 (Li, 2018) and paftools (https://github.com/lh3/minimap2/tree/master/misc) are integrated and applied to call variants from haploid assemblies. For haploid contig alignments and variant selection, mapping quality (>=*20*) is chosen to produce variant candidates in each haplotype. Finally, to generate SNPs, if both haplotypes cover the alternate allele, it is defined as homozygous; if one haplotype covers the reference allele and the other haplotype covers the alternate allele, it is defined as heterozygous.

To generate small Indels and SVs, variant candidates from each haplotype are compared against each other to infer zygosity. For a deletion, if two variant candidates from each haplotype overlap it is called as homozygous, and if the variant candidate only exists in one haplotype (the other haplotype being identical to the reference sequence) it is called as heterozygous. For homozygous insertions, the breakpoints of two variant candidates from each haplotype are within a *20bp* shift tolerance (-*10bp* - *10bp* around breakpoint), and the variant candidate only existing in one haplotype (the other haplotype covered by reference sequence) is called as heterozygous.

Phasing inference

The initial phased SNPs from the Haplotyping module provide the scaffold on which all other heterozygous variants that are discovered by the Variation module are phased (Supplementary Figure 7B). For example, consider the case of one assembled SNP in a phase block, “G|A” or “1|0” (where “A” is the reference allele, and “G” is the alternate allele), and the other neighboring assembled SNP in the same phase block, “C|T” or “0|1” (where “C” is the reference allele, and “T” is the alternate allele): these two phased SNPs have the same genotype and phase in a phase block from the haplotyping module. Therefore, Aquila places the SV into the haplotype that is in the same phase. This is done for all heterozygous variants discovered by the Variation module.

Validation and evaluation methods for variations

To validate the SNP calls from the assembled contigs, we used the GiaB benchmark call sets for both NA12878 and NA24385. They included 3,084,732 SNPs (1,0210,585 homozygous and 1,874,147 heterozygous) for NA12878, and 3,076,552 SNPs (1,180,678 homozygous and 1,895,874 heterozygous) for NA24385. The benchmark allowed us to calculate both the sensitivity and the genotype accuracy of each SNP called from assemblies. We performed the same analysis on the small indel calls with the small indel callset 0.6 from GiaB (NA12878: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/latest/GRCh38/, NA24385: ftp://ftp trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/latest/GRCh38/). For SV validation we applied svviz2 (https://github.com/nspies/svviz2) with PacBio reads (NA12878: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/ and NA24385: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/PacBio_MtSinai_NIST/).

We classified SVs into three categories: Alu, Tandem Repeats, and Other. We selected SVs from 250-350bp for classification as Alu elements by alignment to the AluY consensus sequence, and only SVs that matched at 80% or more were labelled as Alus. We used Tandem Repeats Finder (trf, (Benson, 1999)) to label repeats/non-repeats for all non-Alu-SVs, and the repeat percentage was calculated. Sequences that did not meet the Alu alignment criterion and did not contain a tandem repeat were labeled "Other".

For assemblies of different libraries, the different assemblies could identify different SV candidates. To analyze the overlapping and unique SV calls of three libraries of the same individual, we also merged each SV from different libraries. We applied the same merging criteria when we called homozygous/heterozygous SV from two haplotypes. If the coordinates of two deletions overlapped we defined them to be the same deletion, and if the breakpoints of two insertions were within *20bp*, we defined them to be same insertion.

Ancestral analysis

For each called SV, we extracted the left and right flanking 500bp, and aligned them to the Orangutan and Chimpanzee references. For deletion calls, if the end coordinate of the left flanking sequence was within 2bp of the start coordinate of the right flanking sequence, this SV was defined as “Ins_Ref” (insertion on reference - an actual insertion). If the distance between the end coordinate of the left flanking sequence and the start coordinate of the right flanking sequence was the approximate length of the called deletion (*0.9*SV_size - 1.1* SV_size*), this SV was defined as “Del_Tar” (deletion on target - an actual deletion). For insertion calls, if the end coordinate of the left flanking sequence was within 2bp of the start coordinate of the right flanking sequence, this SV was defined as “Ins_Tar” (insertion on target - an actual insertion). If the distance between the end coordinate of the left flanking sequence and the start coordinate of the right flanking sequence was the approximate length of the called deletion (0.9*SV_size - 1.1* SV_size), this SV was defined as “Del_Ref” (deletion on reference - an actual deletion).

We performed multiple sequence alignments (for each SV and its aforementioned flanking regions annotated as “Ins_Ref”, “Del_Tar”, “Ins_Tar” and “Del_Ref”) with Muscle (Edgar, 2004).

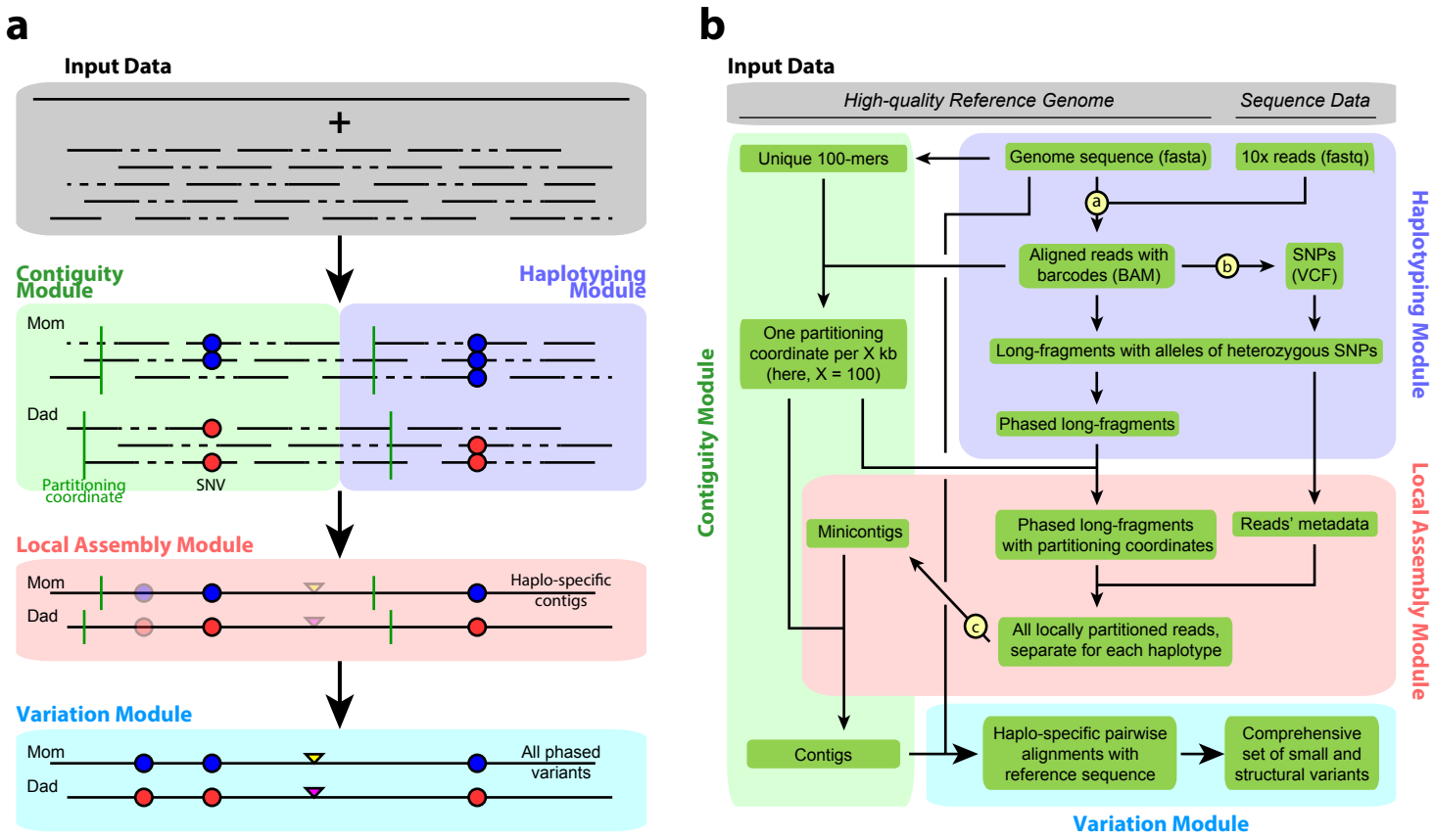

a
Input Data
+
Contiguity Module
Haplotyping Module
Mom
Dad
Partitioning coordinate
SNV
Local Assembly Module
Mom
Dad
Haplo-specific contigs
Variation Module
Mom
Dad
All phased variants
b
Input Data
High-quality Reference Genome
Sequence Data
Unique 100-mers
Genome sequence (fasta)
10x reads (fastq)
a
Aligned reads with barcodes (BAM)
b
SNPs (VCF)
One partitioning coordinate per X kb (here, X = 100)
Long-fragments with alleles of heterozygous SNPs
Phased long-fragments
Contiguity Module
Haplotyping Module
Minicontigs
Phased long-fragments with partitioning coordinates
Reads' metadata
c
All locally partitioned reads, separate for each haplotype
Local Assembly Module
Contigs
Haplo-specific pairwise alignments with reference sequence
Comprehensive set of small and structural variants
Variation Module

Fig 1

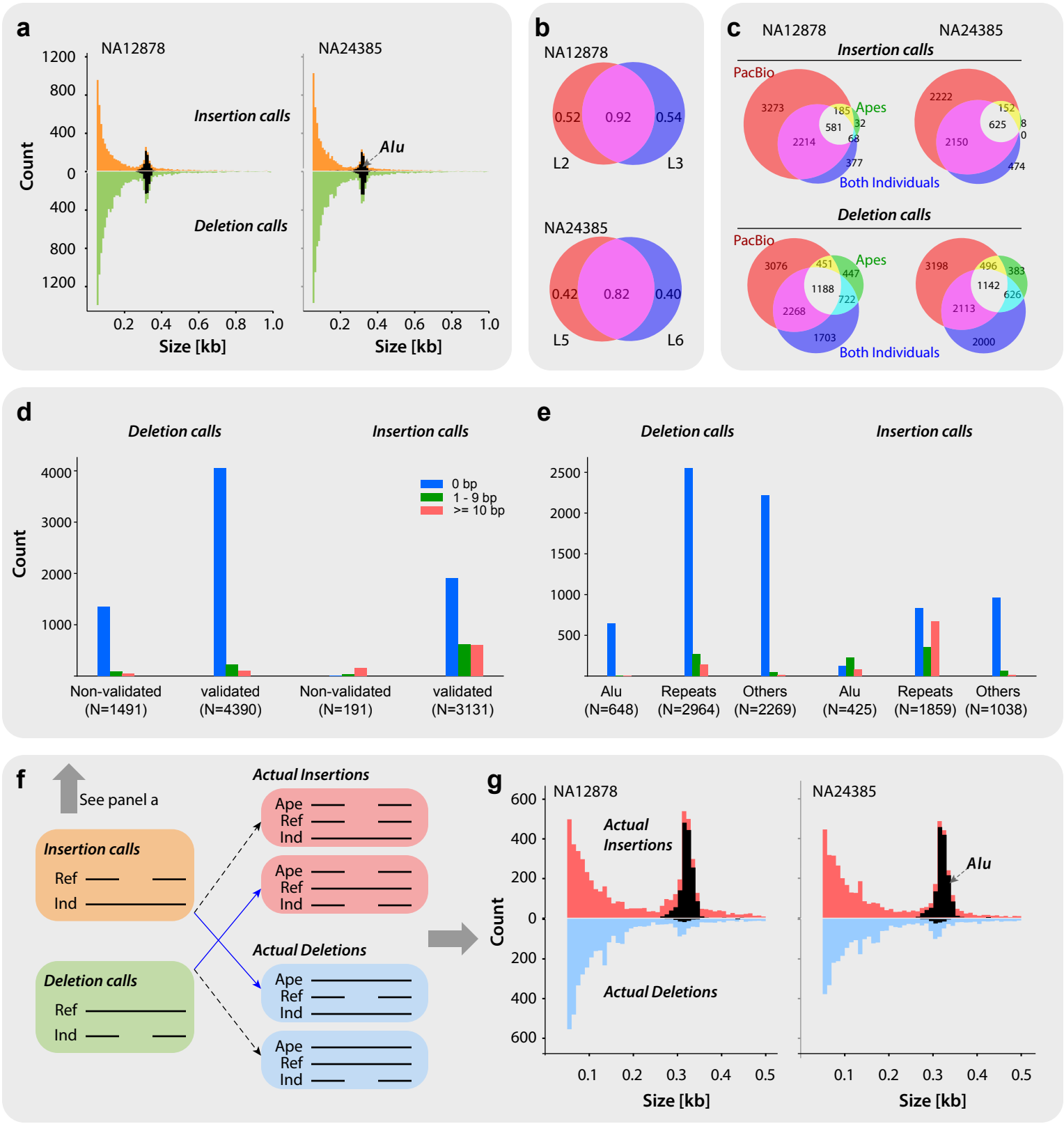

a
NA12878
NA24385
Insertion calls
Deletion calls
Alu
Count
Size [kb]
b
NA12878
0.52
0.92
0.54
L2
L3
NA24385
0.42
0.82
0.40
L5
L6
c
NA12878
NA24385
Insertion calls
PacBio
Apes
Both Individuals
Deletion calls
d
Deletion calls
Insertion calls
0 bp
1 - 9 bp
>= 10 bp
Non-validated (N=1491)
validated (N=4390)
Non-validated (N=191)
validated (N=3131)
e
Alu (N=648)
Repeats (N=2964)
Others (N=2269)
Alu (N=425)
Repeats (N=1859)
Others (N=1038)
f
See panel a
Actual Insertions
Insertion calls
Ref
Ind
Ape
Actual Deletions
Deletion calls
g
Actual Insertions
Actual Deletions
Alu

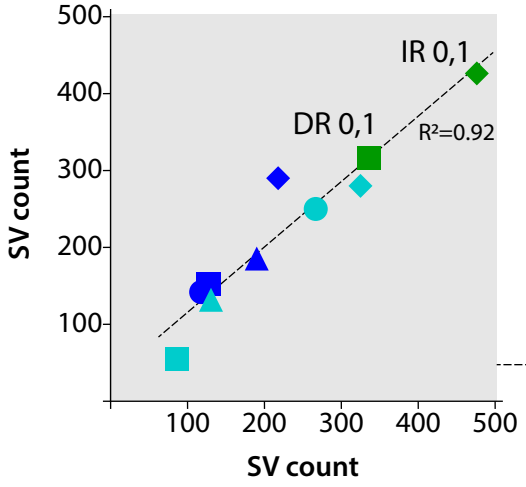
Fig 2

a

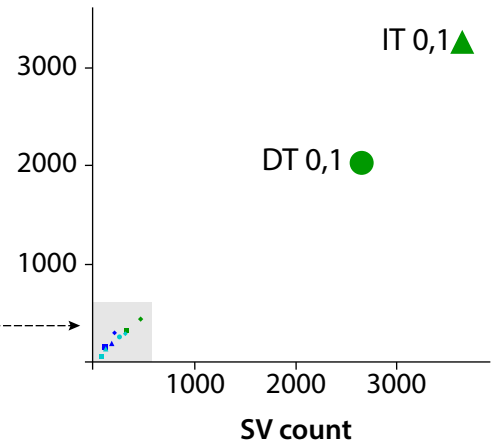
A = Ancestral allele
D = Der = Derived allele

Reference	A	A	A	A	A	A	A	A	A	D	D	D	D	D	D	D	D	D
NA12878-1	A	A	D	D	A	A	A	D	A	A	D	A	A	A	A	D	A	D
NA12878-2	A	A	D	D	A	D	D	D	D	A	D	A	A	D	D	D	D	D
NA24385-1	A	D	A	D	A	A	D	A	A	D	A	A	A	A	D	A	A	D
NA24385-2	A	D	A	D	D	A	D	D	D	D	A	A	D	A	D	D	D	D
Number of:																		
Permutations	N/A	1	1	1	2	2	2	2	4	1	1	1	2	2	2	2	4	N/A
Der alleles	N/A	2	2	4	1	1	3	3	2	3	3	1	2	2	4	4	3	N/A
Class if:																		
Der = Deletion		● DT 0,2			● DT 0,1		● DT 1,2			■ DR 0,2			■ DR 0,1		■ DR 1,2			
Der = Insertion		▲ IT 0,2			▲ IT 0,1		▲ IT 1,2			◆ IR 0,2			◆ IR 0,1		◆ IR 1,2			

b



c



d

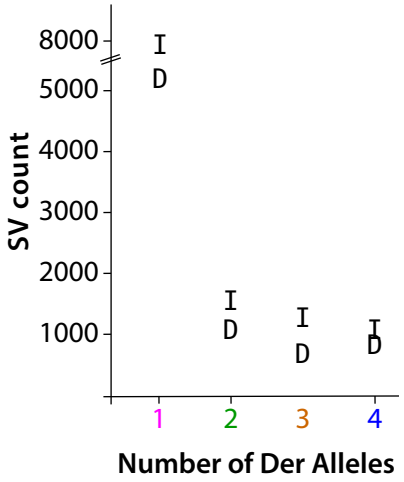


Fig 3

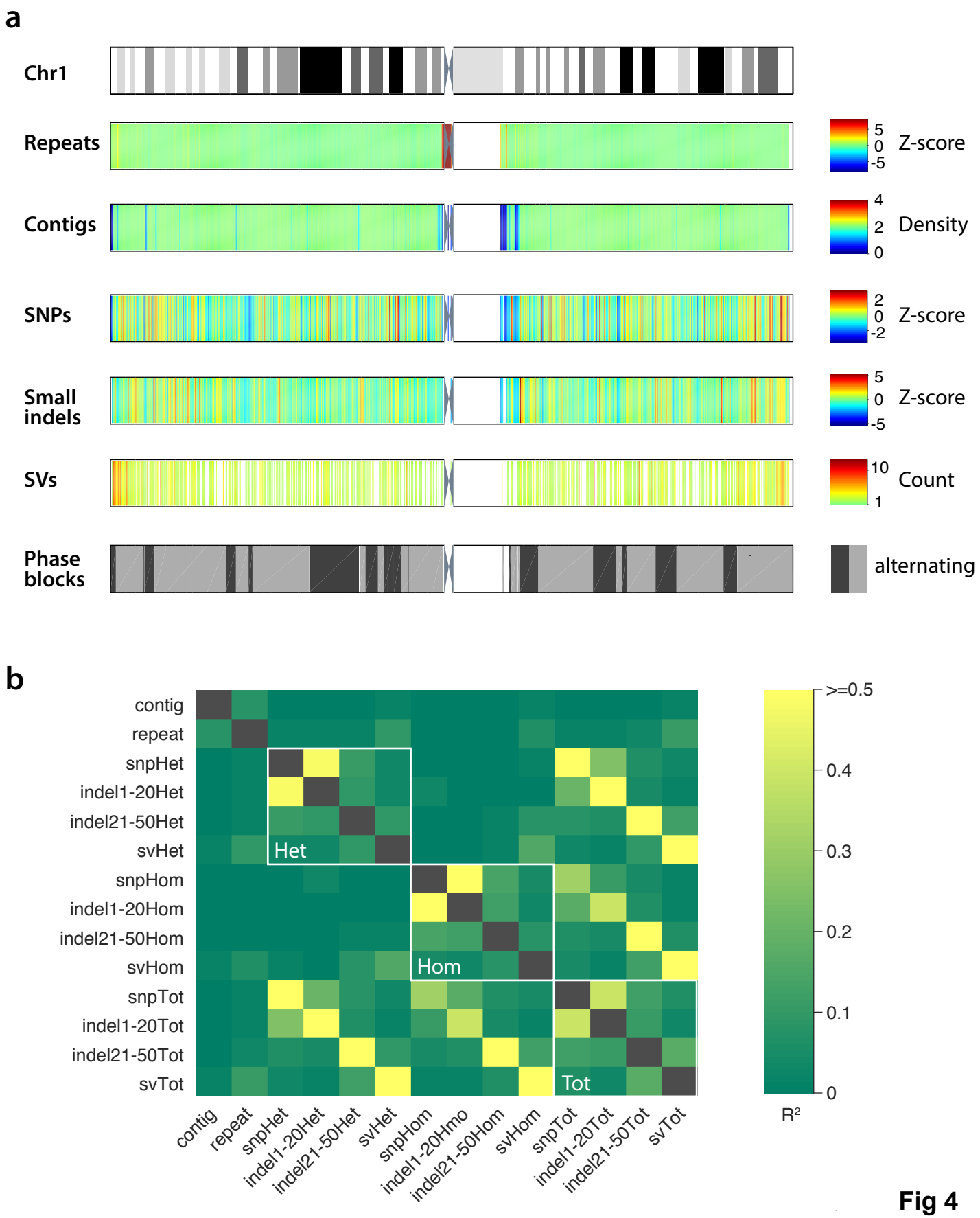

a
Chr1
Repeats
Contigs
SNPs
Small indels
SVs
Phase blocks
5
0
-5
Z-score
4
2
0
Density
2
0
-2
Z-score
5
0
-5
Z-score
10
1
Count
alternating
b
contig
repeat
snpHet
indel1-20Het
indel21-50Het
svHet
snpHom
indel1-20Hom
indel21-50Hom
svHom
snpTot
indel1-20Tot
indel21-50Tot
svTot
Het
Hom
Tot
contig
repeat
snpHet
indel1-20Het
indel21-50Het
svHet
snpHom
indel1-20Hmo
indel21-50Hom
svHom
snpTot
indel1-20Tot
indel21-50Tot
svTot
>=0.5
0.4
0.3
0.2
0.1
0
R²

Fig 4