De novo assembly of haplotype-resolved genomes with trio binning

Sergey Koren1,8, Arang Rhie1,8, Brian P Walenz1, Alexander T Dilthey1,2, Derek M Bickhart3, Sarah B Kingan4, Stefan Hiendleder5,6, John L Williams5,8, Timothy P L Smith7 & Adam M Phillippy1

Complex allelic variation hampers the assembly of haplotype-resolved sequences from diploid genomes. We developed trio binning, an approach that simplifies haplotype assembly by resolving allelic variation before assembly. In contrast with prior approaches, the effectiveness of our method improved with increasing heterozygosity. Trio binning uses short reads from two parental genomes to first partition long reads from an offspring into haplotype-specific sets. Each haplotype is then assembled independently, resulting in a complete diploid reconstruction. We used trio binning to recover both haplotypes of a diploid human genome and identified complex structural variants missed by alternative approaches. We sequenced an F1 cross between the cattle subspecies Bos taurus taurus and Bos taurus indicus and completely assembled both parental haplotypes with NG50 haplotig sizes of >20 Mb and 99.998% accuracy, surpassing the quality of current cattle reference genomes. We suggest that trio binning improves diploid genome assembly and will facilitate new studies of haplotype variation and inheritance.

Genome sequences must be reconstructed from many shorter read sequences in a complex process known as assembly1. Repetitive sequences longer than the sequencing read lengths prevent a complete reconstruction of chromosomes, so assembly typically results in a collection of contiguous sequences (contigs) that are interrupted by repeats or gaps. The advent of long-read sequencing technologies has improved the quality of genome assemblies by resolving many such repeats2. However, even these technologies have not overcome the challenge of completely assembling both haplotypes of a diploid genome. Instead, most genome assembly tools simply co-assemble the haplotypes into a mosaic consensus, resulting in an assembly that does not accurately represent either of the original haplotypes. Collapsing haplotypes into a single consensus representation introduces false variants that are not present in either haplotype, leading to annotation and analysis errors3. Ideally, a genome should be represented as a complete set of haplotypes rather than as an artificial mixture.

A common approach to skirt the issue is to reduce the problem of haplotype variation by sequencing an inbred individual (for example, fly4 or mouse5). However, this is impractical for many species and, even when possible, can result in a genome that is not representative of variation found in the natural population. An alternative approach is to use haploid, clone-based genomic libraries, as was done for the human genome project6. More recently, a diploid human assembly was constructed using tiled fosmids7, but such cloning is often impractical. Alternatively, homozygous cell lines such as CHM1hTERT can be targeted8-10, but such cell lines often develop unstable karyotypes in culture and are not always available. Other attempts have been made to separate haplotypes de novo from whole-genome sequencing. For example, the highly polymorphic sea squirt Ciona savignyi was first assembled using modifications to the Arachne assembler11 designed to split haplotypes on the basis of read overlap information12. However, this was an extreme case, as the reference individual had an estimated heterozygosity of 4.6%. Early attempts to assemble a diploid human genome, with heterozygosity of just 0.1%, first collapsed the haplotypes into a combined assembly and then phased alleles over a short range using pairs of heterozygous variants observed on a single read or read pair13. Current phasing tools operate similarly, and map sequencing reads to a reference sequence to infer blocks of variants that originate from the same haplotype14-16. More sophisticated library preparations, such as chromosome sorting17, Strand-seq18 and Hi-C19, can link variants over a much longer range, delivering chromosome-scale phase blocks. However, methods that rely on reference mapping typically fail in regions of high heterozygosity and/or substantial structural variation between haplotypes, yielding a limited view of genetic diversity.

A more comprehensive solution to the diploid assembly problem is to integrate haplotype separation into the assembly process itself. However, this approach is limited by the fragment length of the sequencing process. Sequencing reads alone do not always contain enough information to link variants across longer regions of homozygosity, resulting in relatively short phase blocks. As a compromise, diploid assemblers such as FALCON-Unzip20 and Supernova21 output
‘pseudo-haplotypes’ that represent a single allele at each position, but do not preserve phase across long homozygous alleles or assembly gaps. In addition, these assemblers can confuse repeats with diverged alleles, leading to artefactual duplications or deletions. One potential solution is to combine long-read sequencing with additional types of information, such as linked reads and/or bacterial artificial chromosomes (BACs)22, Strand-seq23, or Hi-C24. However, none of the de novo assemblers currently integrate these data types, so this can be a manual and expensive process.

We provide a simple solution to the diploid assembly problem that assembles accurate, genome-scale haplotypes de novo. Unlike other methods that are limited to phasing individual chromosomes, our method produces two complete, haploid genomes: one for each parental haplotype. These complete haplotypes are versatile and can be analyzed individually or recombined via alignment into a diploid genome graph. If contiguity is paramount, a maximally contiguous pseudo-haplotype could be defined as a walk through this graph (for example, a repeat might be resolved on one haplotype but not the other). Alternatively, to capture complex structural variation, multiple haplotypes from a population could be combined to create a pan-genome reference. These potential applications are simplified by the initial reconstruction of complete, linear haplotypes.

The separation of haplotypes prior to assembly using a father-mother-offspring trio is the key to our method. Each haplotype is then assembled separately without the interference of inter-haplotype variation. Trios have long been used in genomics to infer inheritance, including for the HapMap project25, the 1000 Genomes Project26 and the creation of ‘platinum’ variant catalogs27. Trios have also been used by trio-sga to simplify heterozygous diploid genome assembly28, but reliance on short-read sequencing limited the haplotype-specific contigs (haplotigs) to an average size of a few kilobases. In contrast, our long-read method enabled the assembly of multi-megabase haplotigs and complete parental haplotypes.

We developed trio binning and found that it can accurately and completely reconstruct parental haplotypes for a wide range of zygosity and genome sizes. We first analyzed benchmark data sets with both high (Arabidopsis) and low (human) levels of heterozygosity, for which prior methods do not completely recover both haplotypes of a diploid genome. We then analyzed a complete diploid assembly of an F1 hybrid between Bos taurus taurus and Bos taurus indicus, and found that the quality of each haplotype exceeded that of even the best livestock reference genomes. These results demonstrate that trio binning of outbred genomes is an easy, accurate and superior method for assembling diploid reference genomes.

RESULTS

Complete haplotype assembly with TrioCanu

We implemented trio binning and haplotype assembly as the TrioCanu module of the Canu assembler29. Our method requires moderate coverage of short, high-quality sequencing reads (for example, 30x Illumina) from two parental genomes to identify short, length k sub-sequences (k-mers) that are specific to each parent. These k-mers are presumed to be specific to the corresponding haplotypes of the offspring. Long reads are then collected from the offspring to sufficiently cover both haplotypes (for example, 40x PacBio per haplotype). Finally, long reads from the offspring are binned into paternal and maternal groups on the basis of the presence of the haplotype-specific k-mers and assembled separately (Fig. 1 and Online Methods).

Trio binning performed best for a uniformly heterozygous offspring, which maximized the probability that any given read would contain at least one haplotype-specific k-mer. Each heterozygous single-nucleotide variant was expected to induce 2k haplotype-specific k-mers. As a result, the fraction of haplotype-specific k-mers was greater than the single-nucleotide heterozygosity. In humans, for example, where single-nucleotide heterozygosity is estimated to be only ~0.1%, nearly 2% of the 21-mers were haplotype specific (Table 1 and Supplementary Table 1). Thus, k-mers are powerful haplotype markers that can also capture complex insertions, deletions and fusion events.

Read classification accuracy depends not only on the zygosity of the offspring, but also on sequencing read length and error rate (Fig. 2). Given the high error rates in current long-read technologies, the k-mer size is also important. It must be long enough to be unique in the genome but short enough that it will not be corrupted by sequencing errors (for example, k = 21 for a 3-Gb genome). Given current long-read sequencing characteristics (read N50 >15 kb and read accuracy >85%), it is possible to bin and assemble nearly all of a human genome. A small fraction of reads will remain unclassified, but in the three data sets that we analyzed, these reads were typically short and were derived from either homozygous alleles or identically heterozygous alleles (that is, both parents shared the same heterozygous genotype). The former reads, being homozygous, could be reassembled with both haplotype bins, whereas the latter were a limitation of trios and would require additional linkage information to be assigned correctly. However, current read lengths typically exceeded the size of such alleles and unclassifiable reads were rare in practice.

Validation on an Arabidopsis cross

The published description of FALCON-Unzip provides a valuable data set for benchmarking diploid assembly algorithms20. The authors crossed two well-characterized strains of Arabidopsis thaliana, Col-0 and Cvi-0, and generated both long-read PacBio and short-read Illumina sequencing reads for the F1 hybrid. Because the parental strains were both highly inbred, recombination was inconsequential and the F1 haplotypes were expected to match the parental genomes, providing a truth set for validation. No short-read data were available for the parental lines, so we inferred haplotype-specific k-mers directly from the assemblies. The heterozygosity was estimated to be 1.36%30, or one variant every 73 bases, representing a best-case scenario for diploid assembly.

TrioCanu successfully classified the A. thaliana F1 reads by haplotype, resulting in unimodal k-mer distributions for the read bins and an assembly that fully resolved both parental haplotypes (Fig. 3, Supplementary Note 1 and Supplementary Fig. 1). In contrast, rather than reporting complete haplotypes, FALCON-Unzip produced pseudo-haplotypes (primary contigs) along with a set of alternate alleles (associated haplotigs) that represent haplotype variants. Thus, homozygous alleles were only represented once and there could be considerable haplotype switching in the pseudo-haplotypes. To compare these results with the TrioCanu haplotigs, we aligned the FALCON-Unzip primary contigs and associated haplotigs to the parental genomes and inferred the correct mapping between assembly and haplotype.

The TrioCanu haplotigs covered 99.50 and 99.00% of the Col-0 and Cvi-0 parental genomes, respectively, and the alignment identity for both was 99.97%. This exceeds the FALCON-Unzip result of 98.47 and 98.53% coverage and 99.94 and 99.92% identity. The NG50 size of the TrioCanu F1 haplotigs was 7.0 and 5.6 Mb for the Col-0 and Cvi-0 haplotypes, respectively, as compared with 7.4 and 6.1 Mb for the inbred parental genomes (NG50 such that 50% of the haploid genome was contained in haplotigs of this size or greater). The NG50 for each haplotype of the diploid TrioCanu assembly
Supplementary Table 2). In comparison, the NG50 of the FALCON-Unzip pseudo-haplotype was slightly longer (8.0 Mb), but contained substantial haplotype switching (Fig. 3d). Compared with either TrioCanu haplotype, the FALCON-Unzip pseudo-haplotype also contained fourfold more duplicated BUSCO genes, and the associated haplotigs were missing fourfold more BUSCO genes (Supplementary Table 2), suggesting that some alternate alleles were incorrectly included in the FALCON-Unzip primary contig set.

Although convenient for some applications, pseudo-haplotypes must be split before reconstructing complete haplotypes with the assistance of additional linkage information. This is a difficult and error-prone process that will propagate any errors in the pseudo-haplotype (for example, collapsed heterozygosity or erroneous duplications). In contrast, TrioCanu haplotigs are inherently partitioned by haplotype and, if desired, can be recombined into a pseudo-haplotype after assembly for greater contiguity. This was simulated by aligning the assemblies of both F1 haplotypes to the *A. thaliana* Col-0 reference and merging alternative haplotigs found to overlap on the reference by >10 kb, which effectively defined alignment paths through the diploid genome graph. Given that this process is alignment based, the corresponding pseudo-haplotype NG50 metric is useful for measuring both the contiguity and structural accuracy of the assembly (NG50 such that 50% of the reference genome is covered by continuous graph alignments of this size or greater). By this measure, the TrioCanu assembly scored better than FALCON-Unzip (NG50 7.1 Mb versus 6.7 Mb), illustrating that the larger FALCON-Unzip contigs were often a result of the fact that it reports a pseudo-haplotype rather than completely resolved haplotypes (Supplementary Table 3).

Accurate structural variant detection from complete haplotypes simply requires a whole-genome alignment of the two. To demonstrate this approach and the accuracy of the TrioCanu haplotypes, we used Nucmer and Assemblytics to identify a total of 4,828 structural variants (SVs) between the Col-0 reference genome (TAIR10) and a *de novo* assembly of the Cvi-0 genome. Assemblytics classified SVs in the range of ≥50 bp and <10 kb. These SVs were then compared with those identified by aligning the TrioCanu and FALCON-Unzip F1 assemblies against the Col-0 reference genome. The positive predictive value (PPV) and sensitivity for the TrioCanu Cvi-0 haplotype assembly representations are shown with homozygous alleles in black and heterozygous alleles (bubbles) colored by haplotype. Graphical representations typically collapse homozygous alleles into a single sequence. A pseudo-haplotype is a path through the diploid graph that separates heterozygous alleles, but does not preserve phase between loci. Complete haplotypes represent all alleles and preserve phase across the entire genome. The ability to assign sequencing reads to a haplotype depends on the zygosity of the genome, the sequencing read length and the sequencing error rate. (b) Log-log plot of minimum required read length (y axis) such that there is a 99% probability of observing at least one haplotype-specific 21-mer per read (negative binomial distribution, Online Methods), dependent on the sequencing error rate (labels) and fraction of haplotype-specific 21-mers in the genome (x axis). Dotted vertical lines mark the fraction of heterozygous 21-mers for *H. sapiens* and the *B. taurus* F1 cross.

---

**Figure 1** Outline of trio binning and haplotype assembly. (a) Two parents constitute four haplotypes, including shared sequences in both parents (solid lines) and sequences unique to one parent (dashed lines). The offspring inherits a recombinant haplotype from each parent (blue, paternal; red, maternal). (b) Short-read sequencing of the parents identifies unique length-k subsequences (k-mers), which can be used to infer the origin of heterozygous alleles in the offspring’s diploid genome. (c) Trio binning simplifies assembly by first partitioning long reads from the offspring into paternal and maternal sets on the basis of these k-mers. Each haplotype is then assembled separately without the interference of heterozygous variants. Unassignable reads are homozygous and can be assigned to both sets or assembled separately. (d) The resulting assemblies represent genome-scale haplotypes and accurately recover both point and structural variation.

---

**Figure 2** Effect of data characteristics on trio binning. (a) Diploid assembly representations are shown with homozygous alleles in black and heterozygous alleles (bubbles) colored by haplotype. Graphical representations typically collapse homozygous alleles into a single sequence. A pseudo-haplotype is a path through the diploid graph that separates heterozygous alleles, but does not preserve phase between loci. Complete haplotypes represent all alleles and preserve phase across the entire genome. The ability to assign sequencing reads to a haplotype depends on the zygosity of the genome, the sequencing read length and the sequencing error rate. (b) Log-log plot of minimum required read length (y axis) such that there is a 99% probability of observing at least one haplotype-specific 21-mer per read (negative binomial distribution, Online Methods), dependent on the sequencing error rate (labels) and fraction of haplotype-specific 21-mers in the genome (x axis). Dotted vertical lines mark the fraction of heterozygous 21-mers for *H. sapiens* and the *B. taurus* F1 cross.
were 99.1 and 99.2%, respectively, as compared with 96.99 and 98.81% for the FALCON-Unzip assembly (primary contigs plus associated haplotigs). However, the FALCON-Unzip PPV was artificially high in this case because variants were only being discovered on one haplotype (that is, the other haplotype matched the reference and induced no variants). As expected, the TrioCanu F1 Col-0 haplotype showed good agreement with the Col-0 reference genome, differing on average by less than 2 variants per 10,000 bases and 108 SVs, which could represent errors in the assembly, errors in the reference or true intraspecies variation.

A personal, diploid human genome

We next evaluated trio binning on a human trio of European descent (father, NA12891; mother, NA12892; daughter, NA12878)22, and compared against Supernova 10x Genomics (linked reads)21 and FALCON-Unzip (PacBio) assemblies of NA12878. As a result of historical population bottlenecks, human genomes typically have a heterozygosity of ~0.1%, which was confirmed for NA12878 via k-mer analysis (Supplementary Fig. 2). This presents a challenge for haplotype recovery, as heterozygous variants are sparse (1 per 1,000 bases on average) and long-range linking information is required to preserve phase. A trio-based approach overcomes this problem because variants can be associated with the parent from which they were inherited, preserving phase across the entire genome.

A TrioCanu assembly of NA12878 from 72× PacBio coverage produced a haplotype NG50 of 1.2-Mb and separate 2.7-Gb assemblies for each parental haplotype (Table 1 and Supplementary Note 1). By comparison, the Supernova assembly from 41× linked-read coverage had a smaller contig NG50 of 103 kb and phase block NG50 of 4.2 Mb. The FALCON-Unzip pseudo-haplotype had a larger contig NG50 of 8.7 Mb and a shorter phase block NG50 of 0.4 Mb. TrioCanu and FALCON-Unzip pseudo-haplotype NG50 sizes were 3.0 and 4.2 Mb, respectively (Supplementary Table 3). Because TrioCanu generated complete haplotypes, the entire genome was in phase and all haplotigs were assigned to the parent from which they were inherited (Supplementary Fig. 3). For example, the TrioCanu paternal haplotype correctly assembled a known CYP2C19 substitution27. Comparing the two TrioCanu NA12878 haplotypes using Assemblytics yielded 6,674 SVs affecting 3.4 Mb of the genome, including 12 inversions with an average size of 19 kb. The alignment included 2.67 million single-nucleotide substitutions, matching the expected heterozygosity. Insertions and deletions (indels) between the haplotypes were well balanced, with an enrichment for 300-bp and 6-kb events, corresponding to human Alu and LINE elements, respectively (Fig. 4a).

To measure accuracy, we compared individual single-nucleotide polymorphisms (SNPs) extracted from the TrioCanu and Supernova assemblies against a gold standard variant call set for NA12878 (ref. 27). Considering only genomic positions covered in both assemblies, the sensitivity of TrioCanu was 91.2 versus 90.9% for Supernova and the PPV was 90.2 versus 93.4%. FALCON-Unzip had lower sensitivity (73.13%) and PPV (70.26%) as a result of incomplete phasing (for example, the associated haplotigs summed to only 65% of the primary assembly length). The slightly lower TrioCanu PPV versus Supernova was likely a result of residual consensus errors in the PacBio assembly. A k-mer analysis also revealed that the TrioCanu assembly was missing some homozygous alleles because of assembly gaps and/or sequencing errors (Supplementary Fig. 4). A higher coverage of long reads, so that each haplotype approaches 30× coverage, could be expected to reduce both consensus errors and missing alleles. Despite the 10x Genomics assembly having longer input fragments than PacBio (mean 51 kb versus 12 kb), the NG50 perfect phase block for Supernova was 4.3 Mb versus 5.6 Mb for TrioCanu. The few TrioCanu phase errors originated from regions in which both parents had identical heterozygous genotypes, which cannot be resolved by the trio method alone without longer read lengths or additional linkage information (Supplementary Note 1).

The TrioCanu assembly was more structurally accurate than the Supernova assembly. In particular, Supernova missed many larger variants, and assembled fewer Alu and LINE indels relative to TrioCanu (Fig. 4a). To better understand the structural accuracy of these assemblies, we examined the major histocompatibility complex (MHC), which is a highly repetitive and heterozygous region of the genome that presents a serious challenge for de novo assembly. This region contains the human leukocyte antigen (HLA) genes, which have been well characterized for NA12878 (ref. 36) and serve as a quality check. Supernova did not accurately assemble either MHC haplotype, failed to capture an HLA-DRB3 gene insertion in the paternal haplotype and incorrectly reported the majority of the MHC class II region as homozygous (Fig. 4b). By comparison, TrioCanu correctly assembled both MHC haplotypes, as demonstrated by perfect HLA typing results and only a single base error in the typing genes. FALCON-Unzip also correctly assembled both MHC haplotypes, but with an additional three errors in the typing genes (Supplementary Tables 4–7 and Supplementary Note 1).

Reference assembly of two cattle breeds using an F1 hybrid

Using trio binning, we sought to generate high-quality, breed-specific reference genomes for Angus and Brahman cattle (examples of the Bos taurus taurus and Bos taurus indicus subspecies, respectively). We collected ~60× Illumina coverage each for an Angus bull and a Brahman cow, and 134× PacBio coverage in reads ≥1 kb for their male F1 offspring. Heterozygosity of the F1 was estimated to be 0.9% (Supplementary Fig. 5). We assigned 98.9% of all F1 bases to a parental haplotype. Unassigned reads were short and not enriched for any particular region of the genome. A separate assembly of these

---

### Table 1 Trio heterozygosity, binning and assembly statistics

| Species     | Total coverage | Haplotype | Haplotype k-mers | Assigned bases | Haplotype coverage | No. haplotigs | Haplotype NG50 (Mb) | Assembly size (Mb) |
|-------------|----------------|-----------|------------------|----------------|---------------------|---------------|---------------------|-------------------|
| A. thaliana | 180.6x         | Col-0     | 12.3%            | 50.8%          | 87.8x               | 215           | 7.03                | 123.52            |
| H. sapiens  | 72.3x          | NA12891   | 0.9%             | 43.2%          | 31.3x               | 7,252         | 1.18                | 2,743.25          |
| B. taurus   | 135x           | Angus     | 2.9%             | 44.2%          | 31.9x               | 7,388         | 1.17                | 2,749.17          |

Total coverage: total F1 sequencing coverage relative to the haploid genome size. Haplotype: F1 parental haplotype. Haplotype k-mers: fraction of diploid genome k-mers specific to this haplotype (16-mers for A. thaliana and 21-mers for H. sapiens and B. taurus). Assigned bases: percentage of all sequencing read bases assigned to this haplotype. Coverage: depth of coverage for all reads assigned to this haplotype. Haplotype NG50: half of the haplotype is contained in haplotigs of this size or larger (based on haploid genome size estimate for each species).
unassigned reads resulted in no contigs over a few thousand base pairs, suggesting that all regions of the genome were successfully partitioned by haplotype.

TrioCanu successfully resolved both F1 haplotypes with a haplotig NG50 exceeding 20 Mb for each (NG50, Angus 26.6 Mb; Brahman, 23.3 Mb; Supplementary Note 1). This far surpasses previous B. taurus taurus\textsuperscript{37} and B. taurus indicus\textsuperscript{38} reference genomes, both of which have contig NG50s <100 kb. The TrioCanu haplotype assemblies were also more contiguous than a Canu assembly of the unbinned data as a result of heterozygous branching in the assembly graph (NG50, 15.6 Mb). A FALCON-Unzip assembly of the combined data achieved an impressive NG50 of 31.4 Mb for its pseudo-haplotype, but with substantial switch error (Supplementary Fig. 6) and a pseudo-haplotype NG50 similar to TrioCanu (4.19 Mb versus 4.20 Mb; Supplementary Table 3). Further analysis of k-mer distributions in the FALCON-Unzip assembly revealed less-complete haplotype separation, with more homozygous k-mers (expected to be 2 copies) occurring either as 1 copy (over-collapsed) or >2 copy (over-split) (Fig. 5a,b and Supplementary Fig. 7). For example, FALCON-Unzip over-collapsed roughly twice as many k-mers as TrioCanu (Supplementary Table 8). Each TrioCanu haplotype was polished using only the haplotype-assigned PacBio reads, and the quality of the final assembly was estimated to be QV47 (accuracy 99.998%; Supplementary Note 1), supporting our contention that higher coverage can overcome the limitations of PPV and missing homozygous alleles observed for the lower-coverage human sample. In addition, polishing only the Brahman haplotype using reads from both haplotypes increased the total number of errors more than twofold, despite the increased coverage, as a result of artifacts introduced by the Angus haplotype. This highlights the advantage of binning reads by haplotype for accurate consensus generation (Supplementary Note 1).

The Angus and Brahman haplotypes aligned to one another with 99.35% identity and contained 25,245 haplotype-specific SVs and 124 inversion breakpoints. Common SV sizes corresponded to known retrotransposon families in the Bos taurus lineage, including the three most common elements: tRNA-Core-RTE (214 bp), RTE-BovB (1,650 bp) and a pseudo-haplotype NGA50 similar to TrioCanu (4.19 Mb versus 4.20 Mb; Supplementary Note 1). The Angus and Brahman haplotypes aligned to one another with 99.35% identity and contained 25,245 haplotype-specific SVs and 124 inversion breakpoints. Common SV sizes corresponded to known retrotransposon families in the Bos taurus lineage, including the three most common elements: tRNA-Core-RTE (214 bp), RTE-BovB (1,650 bp) and a pseudo-haplotype NGA50 similar to TrioCanu (4.19 Mb versus 4.20 Mb; Supplementary Note 1).
and L1 (5,981 bp) (Supplementary Fig. 8). One of the most heterozygous regions between the two haplotypes contained notable copy number variations (CNVs) of GBP2–GBP6 (Fig. 5c). Notably, the Angus haplotype has a large (~140 kb) deletion containing GBP2, whereas the Brahman haplotype includes a complete version of GBP2 transcript variant X8. In addition, GBP6 is partially duplicated only on the Angus haplotype. The FALCON-Unzip assembly was structurally consistent with TrioCanu, but broke this region into nine haplotype-mixed contigs (five primary, four associated) rather than two complete haplotypes (Supplementary Fig. 9). These regions overlapped with a previously reported association of quantitative traits for muscularity and visual conformation scores39, suggesting that our breed-specific haplotypes will be important for understanding growth traits in cattle.

BUSCO34 reported 92.6 and 93.4% complete universal single-copy orthologs and a low rate of duplication (1.0 and 1.1%) for the TrioCanu Angus and Brahman haplotigs, respectively, which is consistent with 93.7% completeness and 1.3% duplication for the current B. taurus taurus Hereford UMD3.1.1 reference26. As with the other genomes, FALCON-Unzip showed more duplicated and fewer complete BUSCO genes (Supplementary Table 2). To further measure the accuracy of the assemblies, we aligned the probe sequence for 735,636 autosomal markers from Illumina’s BovineHD BeadChip to both haplotypes. Only 333 marker loci did not align to either of the TrioCanu haplotypes, and 2,726 and 3,718 were absent from Angus and Brahman, respectively. The 333 marker loci missing from both haplotypes also had low evidence in the parental Illumina data, suggesting that their absence was real in the parental genotypes and not a result of incomplete assembly (Supplementary Fig. 10). The majority of marker sequences missing in one haplotype were also depleted in the corresponding parent’s short read data, but not the other parent, indicating that these are haplotype-specific loci correctly phased by the assembly (Supplementary Figs. 11 and 12). Switch error between the haplotypes was roughly estimated at 0.68% using independent Hi-C data from the F1 (Supplementary Note 1 and Supplementary Table 9).

The Angus and Brahman haplotype assemblies covered 94.2 and 96.2% of the UMD3.1.1 reference genome, respectively, with the Brahman dam haplotype containing the X chromosome and mitochondrial genome (Supplementary Figs. 13 and 14). To our surprise, we identified 3,178 inversions shared by both haplotypes with respect to the reference (mean 9,447 bp, median 4,385 bp;
DISCUSSION

We found that trio binning facilitates complete haplotype assembly for heterozygous diploid genomes, including human. This strategy has several advantages over traditional approaches. First, trio binning recovers the true haplotypes of a viable organism. Both haplotypes of our diploid cattle assembly achieved >20 Mb NG50 haplotype sizes, matching the best contigivities previously reported for homozygous human cell lines sequenced to similar PacBio coverage (for example, CHM1 and CHM13).39 Trio binning is also applicable to organisms that have long generation times or are otherwise recalcitrant to inbreeding. Second, by isolating haplotype variation before assembly, the resulting assembly graphs are simplified. As a result, haplotype-specific assemblies can exceed the continuity of merged haplotype assemblies. After assembly, the resulting haplotypes can be recombined to form a diploid genome graph or contiguous pseudo-haplotype. Third, our approach was able to accurately reconstruct structurally heterozygous alleles that could be important factors in adaptation and immunity (for example, MHC genes) and have previously been linked to quantitative traits (for example, GBP genes). We found that such sequences were often mis-assembled by alternative approaches, and the accurate representation of haplotypes is essential for studies of intraspecific variation, chromosome evolution and allele-specific expression.

We evaluated trio binning on a variety of long-read PacBio coverage, ranging from 70× to 180×. Linked-read assemblies typically require 50× Illumina coverage but do not accurately assemble complex SVs. Standard PacBio assemblies typically require 60× coverage, but are less accurate for identifying small variants, and so are typically combined with Illumina data to maximize base accuracy. In contrast, trio binning accurately identified both SNPs and SVs. We currently recommend a minimum of 40× PacBio coverage per haplotype to achieve accurate consensus sequences, plus an additional 30× Illumina coverage per parent to identify haplotype-specific k-mers. This parental Illumina data can also be used to verify and possibly polish the final assembly. For highly repetitive or less heterozygous genomes, additional long-read coverage may be required to maximize contiguity. Trio binning is compatible with any long-read sequencing technology, such as Oxford Nanopore, and the resulting assemblies will mirror the error characteristics of the chosen platform.

Because trio binning outputs two sets of haplotype-specific reads, it is compatible with any long-read assembler and repeat separation technique for assembling the individual haplotypes. Unlike graph-based assembly representations, which require a specialized biological database.
bioinformatics toolchain, linear haplotypes can be easily analyzed with existing methods. For example, the partitioned read sets can be reused for haplotype-specific gap-filling and consensus polishing, and we found that polishing with haplotype-specific reads achieved a more accurate consensus sequence. Given sufficient haplotype divergence and read lengths, nearly all reads were assigned to the correct haplotype. However, for genomes with lower heterozygosity, long homozygous alleles may receive lower coverage and quality as a result of a lack of assigned reads. In these cases, homozygous reads can be assigned to both haplotypes to boost coverage at the risk of masking some true variants. Additional processing after assembly could correct for this, for example, by mapping the parental short read data to identify missed variants and correct switch error. Alternatively, the accuracy of long-read binning could be improved by more sophisticated classification (for example, using spaced k-mers) or the integration of additional data types (for example, Hi-C). The latter option may allow partial haplotype binning without the use of a trio.

Long-read trio binning, as described here, is the first method able to assemble complete haplotypes from a heterozygous genome and has immediate applications to reference genome construction as well as human and agricultural genomics. New reference genomes will benefit from the improved assembly accuracy and continuity of this approach. For agricultural genomics, trio binning can be used to study breed diversity and has the advantage of producing two reference-quality haplotypes from a single individual. Our assembly of an outbred F1 resulting from a cross between Angus and Brahman cattle produced two breed-specific haplotypes that improved on and corrected the current best reference genomes for both subspecies. These haplotype-specific reference sequences provide an important resource for understanding genetic variation in cattle. The more general idea of haplotype binning should also work well for polyploid plant genomes (for example, bread wheat) by using species markers (rather than parental markers) to pre-partition reads by haplotype. For human genomics, our approach is a viable method for reconstructing complete, personalized haplotypes, and could be used to generate a more complete database of human haplotype variation.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank W. Thompson, K. Kuhn, K. McClure and R. Lee for technical assistance, and T. Graves-Lindsay and Washington University in St. Louis for public release of the PacBio NA12878 data. S.K., A.R., B.P.W. and A.M.P. were supported by the Intramural Research Program of the National Human Genome Research Institute, US National Institutes of Health. S.H. and J.L.W. were funded from the JS Davies bequest to the University of Adelaide. T.P.L.S. was supported by USDA-ARS Project 3040-31000-100-00D. D.M.B. was supported by USDA-ARS Project S909-31000-026-00-D. This research was also supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI17C2988). This work used the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov).

AUTHOR CONTRIBUTIONS

A.M.P. and T.P.L.S. conceived and coordinated the project. S.K. and A.R. designed the trio-binning method. S.K., A.R. and B.P.W. implemented the software. S.K., A.R., B.P.W., A.T.D., D.M.B., S.B.K. and A.M.P. performed analyses. S.H. designed the trio-binning method. S.K., A.R. and B.P.W. implemented the software. S.K., A.M.P. and T.P.L.S. conceived and coordinated the project. S.K. and A.R. designed the NIH HPC Biowulf cluster. Funding (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea. A.R., B.P.W., A.T.D., D.M.B., S.B.K. and A.M.P. performed breeding experiments and sample collections. J.L.W. contributed to development of the concept and provision of samples. T.P.L.S. performed sequencing. S.K., A.R., T.P.L.S., J.L.W. and A.M.P. wrote the manuscript. All of the authors approved the final manuscript.

COMPETING INTERESTS

S.B.K. is a current employee of Pacific Biosciences.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Phillippy, A.M. New advances in sequence assembly. Genome Res. 27, xi–xiii (2017).
2. Koren, S. et al. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol. 14, R101 (2013).
3. Korlach, J. et al. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. Gigascience 6, 1–16 (2017).
4. Myers, E.W. et al. A whole-genome assembly of Drosophila. Science 287, 2196–2204 (2000).
5. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562 (2002).
6. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001).
7. Cao, H. et al. De novo assembly of a haplotype-resolved human genome. Nat. Biotechnol. 33, 617–622 (2015).
8. Steinberg, K.M. et al. Single haplotype assembly of the human genome from a hydatidiform mole. Genome Res. 24, 2066–2076 (2014).
9. Schneider, V.A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 27, 849–864 (2017).
10. Chaisson, M.J. et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature 517, 608–611 (2015).
11. Batzoglou, S. et al. ARCHINE: a whole-genome shotgun assembler, Genome Res. 12, 177–189 (2002).
12. Levy, S. et al. The diploid genome sequence of an individual human. PLoS Biol. 5, e254 (2007).
13. Patterson, M. et al. WhatsApp: weighted haplotype assembly for future-generation sequencing reads. J. Comput. Biol. 22, 498–509 (2015).
14. Ediger, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res. 27, 801–812 (2017).
15. Larkin, D.M. et al. Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. Proc. Natl. Acad. Sci. USA 109, 7693–7698 (2012).
16. Yang, H., Chen, X. & Wong, W.H. Completely phased genome sequencing through chromosome sorting. Proc. Natl. Acad. Sci. USA 108, 12–17 (2011).
17. Falconer, E. & Landsorp, P.M. Strand-seq: a unifying tool for studies of chromosome segregation. Semin. Cell Dev. Biol. 24, 643–652 (2013).
18. Selvaj, S., R Dixon, J., Bansal, V. & Bansal, N. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat. Biotechnol. 31, 1111–1118 (2013).
19. Chin, C.S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods 13, 1050–1054 (2016).
20. Weissenfeld, N.I., Kumar, V., Shah, P., Church, D.M. & Jaffe, D.B. Direct determination of diploid genome sequences. Genome Res. 27, 757–767 (2017).
21. See, J.S. et al. De novo assembly and phasing of a Korean human genome. Nature 538, 243–247 (2016).
22. Porubsky, D. et al. Dense and accurate whole-chromosome haplotyping of individual genomes. Nat. Commun. 8, 1293 (2017).
23. Matthews, B.J. et al. Improved Aedes aegypti mosquito reference genome assembly enables biological discovery and vector control. Nature (in the press).
24. International HapMap Consortium. The International HapMap Project. Nature 426, 789–796 (2003).
25. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65 (2012).
26. Eberle, M.A. et al. Reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Res. 27, 157–164 (2017).
27. Malinsky, M., Simpson, J.T. & Durbin, R. trio-sga: facilitating de novo assembly by highly heterozygous genomes with parent-child trios. bioRxiv Preprint at https://www.biorxiv.org/content/early/2016/05/03/015156 (2016).
28. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27, 722–736 (2017).
29. Vurtue, G.W. et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33, 2202–2204 (2017).
30. Waterhouse, R.M. et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. https://doi.org/10.1093/molbev/msx319 (2017).
31. Salzberg, S.L. et al. GAGE: a critical evaluation of genome assemblies and assembly algorithms. Genome Res. 22, 557–567 (2012).
32. Kurtz, S. et al. Versatile and open software for comparing large genomes. Genome Biol. 5, R12 (2004).
34. Nattestad, M. & Schatz, M.C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021–3023 (2016).
35. Lamesch, P. et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2012).
36. Dilthey, A.T. et al. High-accuracy HLA type inference from whole-genome sequencing data using population reference graphs. *PLoS Comput. Biol.* **12**, e1005151 (2016).
37. Zimin, A.V. et al. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol.* **10**, R42 (2009).
38. Canavez, F.C. et al. Genome sequence and assembly of *Bos indicus*. *J. Hered.* **103**, 342–348 (2012).
39. Zhou, Y. et al. Genome-wide CNV analysis reveals variants associated with growth traits in *Bos indicus*. *BMC Genomics* **17**, 419 (2016).
40. Sedlazeck, F.J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
41. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
42. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
43. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. Assembly of long error-prone reads using repeat graphs. *bioRxiv* Preprint at https://www.biorxiv.org/content/early/2018/01/12/247148 (2018).
44. Chaisson, M.J., Mukherjee, S., Kannan, S. & Eichler, E.E. Duplications de novo using polyploid phasing. in *International Conference on Research in Computational Molecular Biology* (ed. Sahinalp S.) 117–133 (Springer, 2017).
45. English, A.C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**, e47768 (2012).
46. English, A.C. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
47. Loman, N.J., Quick, J. & Simpson, J.T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).
48. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18**, 440–445 (2002).
49. Nattestad, M., Chin, C.-S. & Schatz, M.C. Ribbon: visualizing complex genome alignments and structural variation. *bioRxiv* Preprint at https://www.biorxiv.org/content/early/2016/10/20/082123 (2016).
50. Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B.J. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 574–576 (2017).
ONLINE METHODS

Haplotype k-mer identification. TrioCanu automates k-mer counting, thresholding, and set operations to identify haplotype-specific k-mers. All k-mers are counted using Meryl, a sort-based k-mer counter used within Canu that allows linear-time k-mer set operations. First, a k-mer frequency distribution is obtained by counting k-mers in the parental genomes. This distribution is examined to eliminate k-mers likely to be erroneous (low copy) or from genomic repeats (high copy), leaving only k-mers from unique homoyozygous or heterozygous genome sequences^33. For k-mer coverage x and frequency y, the optimal low coverage threshold is determined by finding the first critical point y = 0 and its corresponding coverage x0 and frequency y0. The same frequency cutoff y0 is used to determine the high coverage threshold x1. A k-mer set D for each haplotype is drawn from all haplotype k-mers. For two parental haplotypes i and j, haplotype-specific k-mer sets are then constructed as H1 = D1 − D5 and H5 = D5 − D1. Classification k-mers with coverage x0D1 < x < x1D5 are selected from H1 and x0D5 < x < x1D1 from H5.

Smaller k-mers are more likely to avoid sequencing error in the reads, so it is preferable to choose a small value for k. However, k must be large enough to minimize random k-mer collisions in the genome. For example, the total space of 16-mers is only 4^16 or 4.29 billion, close to the total number of k-mers in a 3 Gb mammalian genome, increasing the chance that some k-mer may occur multiple times simply by chance (and not homology). Given a genome size G and tolerable collision rate p, an appropriate k can be computed as k = log2(G/p) + 2. According to this formula, we used k = 16 for A. thaliana and k = 21 for H. sapiens and B. taurus. A. thaliana, H. sapiens, and B. taurus were assembled before TrioCanu automation, and the haplotype-specific k-mer thresholds were identified manually as described in Supplementary Note 1. For A. thaliana, which was lacking Illumina data for the parents, k-mers were collected from assemblies of the parents, excluding repetitive k-mers occurring more than 10 times. For H. sapiens and B. taurus, haplotype-specific k-mers were collected from unassembled, short read sequencing of the parents. Low and high k-mer coverage thresholds were chosen manually as x0 = 30 and x1 = 160 for H. sapiens and x0 = 11 and x1 = 100 for B. taurus. Retrospective application of the automated thresholding method selected similar thresholds of ([25,143], [27,147]) and ([10,57], [10,67]) for H. sapiens and B. taurus, respectively.

Haplotype binning. Haplotype binning is a general strategy for partitioning a read set into haplotype groups before assembly. The number of haplotypes is not necessarily limited to two. Given N haplotypes, the goal is to identify haplotype-specific k-mers that are exclusive to one haplotype. Given a database of haplotype-specific k-mers, the number of specific k-mers from each haplotype is counted in each read. It is expected that k-mers in a single read will be from the same haplotype, but due to sequencing errors it is possible to observe spurious k-mers from a different haplotype. Therefore, the observed haplotype-specific k-mer counts are normalized by the database size to control for the different k-mer set sizes of the parents. Reads are then assigned to the haplotype with the most matching haplotype-specific k-mers. In the event of a tie or too few haplotype-specific k-mers, the read is marked as ambiguous. Finally, the N read bins are passed to Canu for assembly, with the option to include the ambiguous reads in all bins.

Whether a read can be correctly classified is a function of the k-mer heterozygosity h, read length l, read error rate e, and k-mer size k. For simplicity of modeling, errors and haplotype differences are assumed to be random point mutations, and heterozygosity h is defined as the fraction of genomic k-mers that are haplotype specific. It is assumed that k is large enough to avoid chance collisions. A read of length l contains l − k + 1 k-mers. The probability of a single k-mer surviving uncorrupted is (1−e)^k, and the expected number of uncorrupted k-mers in a read is l(l + k − 1)(1−e)^k. The expected number of haplotype-specific k-mers in a read is h(l(l + k − 1)(1−e)^k), and the number of surviving haplotype specific k-mers in a read is h(l(l + k − 1)(1−e)^k). Thus, for a typical long sequence read with e = 0.12 and l = 15,000, and k-mer heterozygosity h = 0.001, the expected number of surviving haplotype-specific 16-mers is 2 and 21-mers is 1. Increasing divergence to h = 0.01 increases the expected number of 16-mers to 19 and 21-mers to 10.

Validation. Classification accuracy was evaluated using a truth set of A. thaliana parental reads. The simple majority-wins classification heuristic showed a good sensitivity/specificity tradeoff, exceeding 80% true positive rate (TPR) with <20% false positive rate (Supplementary Fig. 21). We further simulated increased heterozygosity within each parent to measure the effect on k-mer classification. Read classification is more difficult with increasing heterozygosity in the parents, and performance dropped to 74% TPR with <28% false positive rate when parental heterozygosity was increased to 2% (Supplementary Fig. 21). False positives include homoyozygous reads which do not affect the resulting assembly, and a small fraction of mis-classified heterozygous reads. These will be outvoted by the majority of correctly classified reads when building the haplotype consensus. If high specificity is required, the classifier can be tuned to require more than a simple majority of haplotype-specific k-mers.

Assembly alignments were performed with MUMmer 3.23 (ref. 33) with the commands nuclmer -maxmatch -l 100 -c 500 ref.fa asm.fa dnadiff -d out.delta

GRCh38 (ref. 9) excluding ALT loci was used for H. sapiens. TAIR10 was used for A. thaliana. GCF_000803555.6 with chromosome Y from NC_016145.1 was used for B. taurus and AGFL0000000.1 for B. indicus. A genome size of 119,667,750 was used for A. thaliana (TAIR 10 length), 3,098,794,149 for H. sapiens (GRCh38 primary assembly excluding alternates), and 2,713,423,491 for B. taurus (the UMD 3.1.1 reference plus the Y chromosome).

NGA50 statistics for individual assemblies were computed using MUMmer's dnadiff tool. One-to-one alignment intervals for the contigs versus the reference (1coords output) were filtered to only include those intervals >10 kb and ≥97% identity. To ignore small SVs versus the reference, same-strand alignments within 2,000 bp of each other were merged. For TrioCanu and FALCON-Unzip assemblies, this process was repeated for the combined assemblies (all haplotigs from both haplotypes) to compute a pseudo-haplotype NGA50. In this case, same-strand alignments between alternative haplotigs that overlapped by more than 10 kb on the reference were merged to represent a path through the diploid genome graph.

Parent-specific k-mers were used to estimate switch error within assembly contigs. MHC typing was run as previously described^44 with a previously published^31 truth set. B. taurus markers used in BovineHD BeadChip (Illumina Inc., San Diego, CA) were used to identify missing regions in the assemblies as well as haplotype-specific sequences. Illumina data was used to estimate QV by mapping with BWA-MEM^34 and identifying variants with Freebayes^35. Repeats in the Bos taurus genome were downloaded from the UCSC genome browser^36 (Supplementary Note 1).

Sample preparation and sequencing of the cattle trio. The animals used were part of the Davies Epigenetics and Genetics Resource at the University of Adelaide, Australia, and were established and sampled using procedures approved by the animal ethics committee of the University. A 2-year-old cow of the Brahman breed (subspecies Bos taurus indicus) was bred by artificial insemination using semen from a 5-year-old bull of the Angus breed (Bos taurus taurus). The Brahman female had been previously typed for mitochondrial DNA haplotype to verify the maternal lineage as indicus-specific. At day 153 post-insemination, the animal was sacrificed and the fetus removed for dissection. The fetal lung was removed immediately into liquid nitrogen, and DNA was extracted using a salting out procedure. Briefly, approximately 100 mg of tissue was ground under liquid nitrogen to a powder and transferred to a tube containing 2.26 ml of nuclei lysis mixture (2 ml buffer NBF composed of 10 mM Tris-HCl pH 8.0, 0.4 M NaCl, 2 mM EDTA, plus 0.2 ml 10% SDS, plus 0.06 ml 10 mg/ml RNase A). Tissue and solution were mixed by inversion for 2 min, then set to shake slowly at 37 °C 1 h. Protein digestion was performed by adding 0.025 ml Proteinase K (20 mg/ml) and returning to the shaker overnight (approximately 16 h). Protein was removed by addition of 1.25 ml of saturated NaCl, followed by vigorous hand shaking for 15 s and centrifugation 2,250 g, 20 min, 4 °C. The clarified supernatant was transferred to a tube containing 8 ml of cold 100% ethanol, and DNA was precipitated by gentle rocking of the solution. The DNA was transferred using a glass rod and washed twice.
in tubes containing 5 ml of 70% ethanol. The pellet was then transferred to a 1.5 ml tube and air dried for 10 min at 22 °C. DNA was removed from the glass rod by dissolving in 0.25 ml of solution containing 10 mM Tris-HCl pH 8.0 and 0.1 mM EDTA overnight at 4 °C. Parental DNA samples were extracted using standard phenol-chloroform based procedures.

Sequence libraries for the parents and the fetus were prepared with TruSeq PCR-free preparation kits as directed by the manufacturer (Illumina). The three libraries were sequenced in separate runs, with no other libraries present in the flow cell, on a NextSeq500 instrument using 2 × 150 paired end reads with High Output Kit v2 chemistry. The libraries employed unique indexes and, despite being in separate runs, only reads with appropriate indexes for the library were used for analysis to prevent any cross-contamination between the sire, dam, or fetal library data.

Libraries for SMRT sequencing were constructed as recommended by the manufacturer (Procedure P/N 100-286-000-07, Pacific Biosciences), using a 15 kb cutoff for size selection on the BluePippin instrument (Sage Science). A total of 12 library preparations were used, nine of which were sequenced using P6/C4 chemistry on an RSII instrument (Pacific Biosciences) which generated approximately 152 Gb of sequence, and the other three libraries were sequenced on a Sequel instrument which generated another 205 Gb.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Code availability.** Prototype code used to build k-mer sets, subtract parental k-mers, and classify reads is available at [https://github.com/skoren/triobinning](https://github.com/skoren/triobinning). Scripts and as **Supplementary Code**. TrioCanu is implemented as a module of Canu v1.7 and freely available at [https://github.com/marbl/canu](https://github.com/marbl/canu).

**Data availability.** Sequencing data for the cattle trio is available under NCBI BioProject PRJNA432857. All other sequencing data were obtained from public sources. Data accessions, software versions, and commands used to produce the described results are provided in **Supplementary Note 1**. Assembly files, accession numbers and other miscellaneous information can be found at [https://gembox.cbcb.umd.edu/triobinning/index.html](https://gembox.cbcb.umd.edu/triobinning/index.html).

51. Kajitani, R. et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014).
52. Fofanov, Y. et al. How independent are the appearances of n-mers in different genomes? *Bioinformatics* **20**, 2421–2428 (2004).
53. Dilthey, A., Cox, C., Iqbal, Z., Nelson, M.R. & McVean, G. Improved genome inference in the MHC using a population reference graph. *Nat. Genet.* **47**, 682–688 (2015).
54. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at [https://arxiv.org/abs/1303.3997](https://arxiv.org/abs/1303.3997) (2013).
55. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at [https://arxiv.org/abs/1207.3907](https://arxiv.org/abs/1207.3907) (2012).
56. Casper, J. et al. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.* **46**, D762–D769 (2018).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
  - Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
  - For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  - Give P values as exact values whenever suitable.
  - For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
  - For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
  - Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated
  - Clearly defined error bars
    - State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

No software was used.

Data analysis

All software versions and commands are described in Supplementary Note 1. All software used is freely available.

- FALCON-Unzip builds 08/24/2017 and 11/02/2017
- ngmlr 0.2.6
- sniffles 1.0.8
- busco 3.0.2
- bedtools 2.15.0
- nucmer 3.23.0
- freebayes v1.1.0-50-g61527c5
- ArrowGrid github tag 563dcb943160f320fed8bae12cf5b1cf5f586639
- smrtlink 5.1.0.26412
- bax2bam 0.0.8
- pbalign 0.3.1
- variantCaller 2.2.2
- dataset 0.1.27
- canu/meryl github tag c3488e1ab1355da50c60d6ba0b963588def91021
- canu github tag 78d0188e8186476caadfd93062601a38773a952e
Data

Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequencing data for the cattle trio is available under NCBI BioProject PRJNA432857. All other sequencing data was obtained from public sources. Data accessions, software versions, and commands used to produce the described results are provided in the Supplementary Note.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☑ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | N/A. The manuscript describes an algorithm for genome assembly, not a statistical test. |
|-------------|-----------------------------------------------------------------------------------------|
| Data exclusions | N/A. The manuscript describes an algorithm for genome assembly, not a statistical test. |
| Replication | The assembly algorithm was tested on reference plant and human genomes to measure its accuracy and effectiveness. Accuracy of the cattle trio assembly was measured using independent data sources such as Illumina sequencing and known SNP chip markers. |
| Randomization | N/A. The manuscript describes an algorithm for genome assembly, not a statistical test. |
| Blinding | N/A. The manuscript describes an algorithm for genome assembly, not a statistical test. |

Materials & experimental systems

Policy information about availability of materials

☐ Involved in the study
☐ Unique materials
☐ Antibodies
☐ Eukaryotic cell lines
☐ Research animals
☐ Human research participants

Research animals

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Animals/animal-derived materials

As described in the Methods section: "The animals used were part of the Davies Epigenetics and Genetics Resource at the University of Adelaide, Australia, and were established and sampled using procedures approved by the animal ethics committee of the University. A two-year-old cow of the Brahman breed (subspecies Bos taurus indicus) was bred by artificial insemination using semen from a five-year-old bull of the Angus breed (Bos taurus taurus). The Brahman female had been previously typed for mitochondrial DNA haplotype to verify the maternal lineage as indicus-specific. At day 153 post-insemination, the animal was sacrificed and the fetus removed for dissection."

Nature Biotechnology: doi:10.1038/nbt.4277
Method-specific reporting

n/a Involved in the study
☒ ChIP-seq
☒ Flow cytometry
☒ Magnetic resonance imaging