Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies

Arang Rhie[1,†], Brian P. Walenz[1], Sergey Koren[1], Adam M. Phillippy[1]

[1] Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD USA

[†] Corresponding author: A.R. <arang.rhie@nih.gov>

Abstract
Recent long-read assemblies often exceed the quality and completeness of available reference genomes, making validation challenging. Here we present Merqury, a novel tool for reference-free assembly evaluation based on efficient k-mer set operations. By comparing k-mers in a *de novo* assembly to those found in unassembled high-accuracy reads, Merqury estimates base-level accuracy and completeness. For trios, Merqury can also evaluate haplotype-specific accuracy, completeness, phase block continuity, and switch errors. Multiple visualizations, such as k-mer spectrum plots, can be generated for evaluation. We demonstrate on both human and plant genomes that Merqury is a fast and robust method for assembly validation.

Keywords
Genome assembly, assembly validation, benchmarking, k-mers, haplotype phasing, trio binning

Availability of data and material
Project name: Merqury
Project home page: https://github.com/marbl/merqury, https://github.com/marbl/meryl
Archived version: https://github.com/marbl/merqury/releases/tag/v1.0
Operating system(s): Platform independent
Programming language: C++, Java, Perl
Other requirements: gcc 4.8 or higher, java 1.6 or higher
License: Public domain (see https://github.com/marbl/merqury/blob/master/README.license)
Any restrictions to use by non-academics: No restrictions applied

Background

With recent advances in long-read[1–3] and long-range sequencing technologies[4–6], new assembly pipelines are generating more continuous, complete, and accurate diploid genome assemblies than ever before[4,7–14].

However, *de novo* assembled genomes are difficult to validate due to the lack of a known truth. Existing methods use Illumina reads to infer base-level accuracy by aligning the reads to the assembly for evaluation[15]. Base errors in the consensus are detected as variants (SNPs or small indels) when aligning the short reads. However, this method is heavily reliant on the short-read mapping, which could be biased in repetitive regions, under-collapsed regions, or regions of low consensus accuracy. For measuring completeness and false duplications, near-universal single-copy orthologs (BUSCOs)[16] have been widely used to evaluate the gene content of assemblies. BUSCO is robust for species that have been widely studied, such as human and mouse. However, this analysis can be inaccurate when the newly assembled genome contains true copy number or sequence variants that were not considered when building the initial BUSCO gene set. In addition, BUSCO only examines conserved single-copy genes and fails to evaluate the most difficult-to-assemble regions of the genome.

In contrast, k-mers (genomic substrings of length *k*) can be used in a reference-free manner for assessing genome assembly quality metrics. Genome assembly validation via k-mer copy number analysis was introduced by Mapleson *et al.* in their KAT tool[17], which enables visual inspection of k-mer spectra plots to identify artificial duplications and missing sequences. Merqury takes much of its inspiration from the ideas introduced by KAT. QUAST-LG[18] is another assembly validation tool that provides both BUSCO and KAT measurements, as well as alignment against a closely related reference genome. However, many of QUAST's metrics are based on the

reference alignment, which incorrectly reports true variants in the assembled genome as potential mis-assemblies.

Assessing haplotype phasing without a truth set is also challenging. Diploid genome assemblers generate both primary and alternate assemblies representing the two haplotypes. The primary assembly is typically a pseudo-haplotype that captures both the homozygous regions along with a single copy of the heterozygous alleles. Such a pseudo-haplotype does not guarantee long-range phasing, so to estimate phase block statistics, the alternate alleles must be mapped back to the primary assembly to determine regions corresponding to the primary-alternate haplotype phase blocks[9]. However, this can be challenging when the alternate alleles do not map well to the primary sequence due to high sequence divergence or mis-assemblies. Moreover, long-read assemblies often collapse regions of low heterozygosity, which are excluded when calculating phase block statistics, thus over-representing the correctness.

Alternative methods report phasing statistics from small variants (mostly SNPs) called with short-read mapping[8,19–22], or use benchmark genomes that have curated, phased variation call sets[23–26]. Both methods rely on a reference sequence as the primary source to detect heterozygous variations. However, these reference sequences are incomplete, and recent studies have demonstrated the shortcomings of the current human reference genome and variant call sets[8,20,27]. For example, the highly variable major histocompatibility complex (MHC) is excluded from the Genome in a Bottle (GIAB) and Global Alliance for Genomic Health (GA4GH) reference panels[25] due to its repetitive nature and need of a specialized mapping strategy to account for the high allelic diversity[28]. Moreover, reference-guided strategies require significant manual curation and effort and will not scale as large cohort sequencing projects become common[29,30]. This validation strategy is also not applicable to any species without a curated and complete reference[31].

To overcome these limitations, we developed Merqury, which generates assembly assessment metrics using k-mers alone. Merqury compares a set of k-mers derived from unassembled, high accuracy sequencing reads to a genome assembly for evaluation. The generated assembly metrics include consensus quality (QV), k-mer completeness, and visual aids for representing the copy number spectrum and k-mer coverage across the assembly. When parental genomic sequences are available (either assembled or unassembled), Merqury can also report haplotype completeness, phase block statistics (including switch error rates), and visual representations of phase consistency for the child's genome. We show that k-mer-based assembly validation produces comparable or better results than existing methods, such as BUSCO gene completeness and mapping-based measurements.

Results

To demonstrate the ability of Merqury to evaluate the accuracy, completeness, and phasing of an assembly, we first applied it to an *Arabidopsis thaliana* F1 hybrid[20], for which the parental strains (Col-0 and Cvi-0, simplified as Col and Cvi) have also been sequenced. For a comparison of multiple assemblies, we demonstrate Merqury on haplotype-resolved (TrioCanu[10]), pseudo-haplotype (FALCON-Unzip[20]), and mixed-haplotype (Canu[32]) assemblies of this hybrid genome. Total assembly size is typically used as a rough measure of haplotype completeness. For example, the TrioCanu haplotype assemblies have similar total bases, 122~124 Mbp (**Table 1**), close to the expected haploid genome size of 130 Mbp, indicating the haplotype assemblies are well balanced (assuming haplotype specific bases are evenly inherited). In comparison, the primary assembly of FALCON-Unzip has ~35 Mbp bases more than the alternate assembly. However, it is difficult to understand where this difference originates from the assembly size alone. The mixed-haplotype Canu assembly, in comparison, is ~100 Mbp larger than the expected genome size of 130 Mb. Again, we can assume this assembly resolved both haplotypes, but since the haplotypes have been combined in a single assembly we cannot know the composition from

73 the size and continuity measures alone. In the following sections we describe how Merqury's

74 statistics and plots can be used to dissect and understand these assemblies.

Table 1. Merqury quality and completeness statistics for example *Arabidopsis thaliana* and human genome assemblies.

	Assembly	Continuity				QV			k-mer completeness				BUSCO (%)			
		Total	# Contigs	Max	NG50	**Meryl**	**Mash**	Map	**All**	**Hap1**	**Hap2**	**PPV**	Comp.	Dup.	Frag.	Mis.
***A. thaliana* Col x Cvi F1**	**TrioCanu**															
	Col	124	215	13.1	4.6	35.1	34.6	37.5	83.8	99.5	0.7	99.3	98.2	1.6	0.5	1.3
	Cvi	122	163	12.2	5.6	36.4	35.9	38.6	83.6	1.1	99.4	98.9	98.0	1.4	0.4	1.6
	Col + Cvi	246	378	13.1	5.5	35.7	33.0	38.0	98.3	99.7	99.4	99.1	98.2	78.8	0.2	1.6
	FALCON-Unzip															
	pri	140	172	13.3	8.0	34.8	33.9	36.9	87.1	65.2	59.8	N/A	98.1	6.2	0.3	1.6
	alt	105	248	11.6	4.3	38.3	37.9	39.9	74.5	38.2	40.6	N/A	93.1	2.0	0.3	6.6
	pri + alt	245	420	13.3	6.9	36.0	33.5	37.9	97.8	99.1	98.2	N/A	98.1	93.2	0.3	1.6
	Canu	248	2,368	6.9	2.3	29.3	26.8	27.2	95.7	90.5	90.8	N/A	97.6	61.5	0.4	2.0
***H. sapiens* NA12878**	**TrioCanu**															
	mat	2,749	7,388	9.0	1.1	31.3	30.4	34.1	94.0	90.7	0.9	99.1	86.5	0.8	6.9	6.6
	pat	2,743	7,252	11.5	1.1	31.0	30.1	34.1	93.7	1.0	91.0	98.8	85.1	0.7	7.8	7.1
	mat + pat	5,492	14,640	11.5	1.1	31.1	27.5	34.1	98.2	90.9	91.2	99.0	90.0	47.3	5.0	5.0

All bases in the continuity stats are in Mbp. Haploid genome size of 130 Mbp and 3.2 Gbp was used for NG50 in the *A. thaliana* and *H. sapiens* NA12878 haploid assemblies, respectively, with twice the haploid genome size for combined assemblies. Mercury-specific column headers are in bold, including k-mer-based quality (QV) and completeness estimates. Merqury includes both exact (Meryl) and approximate (Mash) methods for measuring k-mer QV, while completeness uses only the exact k-mer counting method. Consensus QV scores are Phred-scaled where $QV = -10 \log_{10} E$ for a probability of error E at each base in the assembly. K-mer completeness is measured by the fraction of all distinct, reliable k-mers (all) and haplotype specific k-mers. Hap1 and Hap2 corresponds to Col and Cvi in *A. thaliana*, maternal and paternal in NA12878. BUSCO v3 was run with embryophyta_odb9 (n=1440) for *A. thaliana* and mammalia_odb9 (n=4104) for NA12878. Metrics generated with Merqury are highlighted in bold. Map. Mapping; Comp., Complete; Dup. Duplicated; Frag. Fragmented; Mis. Missing.

75 ***Copy number spectrum***

76 We counted k-mers from Illumina whole-genome sequencing of the *A. thaliana* F1 hybrid, as
77 well as from each assembly, using Meryl, a k-mer counting tool we extended to support k-mer set
78 operations for Merqury (**Methods**). The copy number spectrum plot, known as "spectra-cn" plot[17]
79 tracks the multiplicity of each k-mer found in the Illumina read set (**Fig. 1a**) and colors it by the
80 number of times it is found in a given assembly (**Fig. 1b**). The result is a set of histograms relating
81 k-mer counts in the read set to their associated counts in the assembly. Here the Illumina dataset
82 (which we will refer to as the "read set") was sequenced to an average coverage of 45x, so we
83 expect a histogram peak near x=45 corresponding to k-mers present in both haplotypes, and a
84 peak at half coverage (x=22) representing k-mers found on only one haplotype (**Fig. 1a,** in
85 practice, these peaks are shifted slightly lower due to the effects of sequencing error and sampling
86 only l-k+1 k-mers for each l-sized read). We refer to these as 2-copy and 1-copy k-mers,
87 respectively, to indicate the number of times they appear in the true genome. Additional peaks
88 may appear for polyploid genomes, but for the remainder of this paper we will assume a diploid
89 genome.

90 Thus, when a k-mer is found approximately 22 times in the *A. thaliana* read set, we expect it
91 to be found only once in the assembly, as it is likely a 1-copy, haplotype-specific (heterozygous)
92 sequence (**Fig. 1b**). In the spectra-cn plot, k-mers are colored based on their count in the
93 assembly. For a complete haplotype-resolved assembly, where even the homozygous part of the
94 genome is included in both haplotypes, we expect most k-mers in the 2-copy peak to be found
95 twice in the assembly (**Fig. 1b**). For partially phased assemblies, 2-copy k-mers may be found
96 either once or twice in the assembly (e.g. **Fig. 2**), depending on which homozygous sequences
97 of the genome were separated and which were collapsed. In contrast, a pseudo-haplotype
98 collapses homozygous alleles, so 2-copy k-mers are expected to appear only once in the
99 assembly. One notable exception is haplotype-specific duplications, which can occur in two

100 copies on the same haplotype, and thus may also appear in two copies in a pseudo-haplotype
101 assembly.

102 Assuming no serious biases in the read set, a clean spectra-cn plot is a necessary, but not
103 sufficient, condition of assembly quality. Quoting KAT author Bernardo Clavijo "if your spectra
104 look right, your assemblies could still be wrong, but if your spectra look wrong your assemblies
105 can't be right.". That is to say, certain structural errors may not be reflected by the spectra, but
106 any conflict with the copy numbers of the read set indicate assembly problems. The k-mers found
107 only in the read set (black) at low frequency are almost always indicative of sequencing error in
108 the read set, however higher-frequency k-mers found only in the read set indicate missing
109 sequences in the assembly (e.g. black k-mers within the 1- or 2-copy peaks). Likewise, any k-
110 mers with a higher copy number in the assembly than would be predicted by the read set are
111 indicative of artificial duplications in the assembly, e.g. see the 2-copy k-mers appearing three
112 times in the Canu assembly shown in **Fig. 1e**.

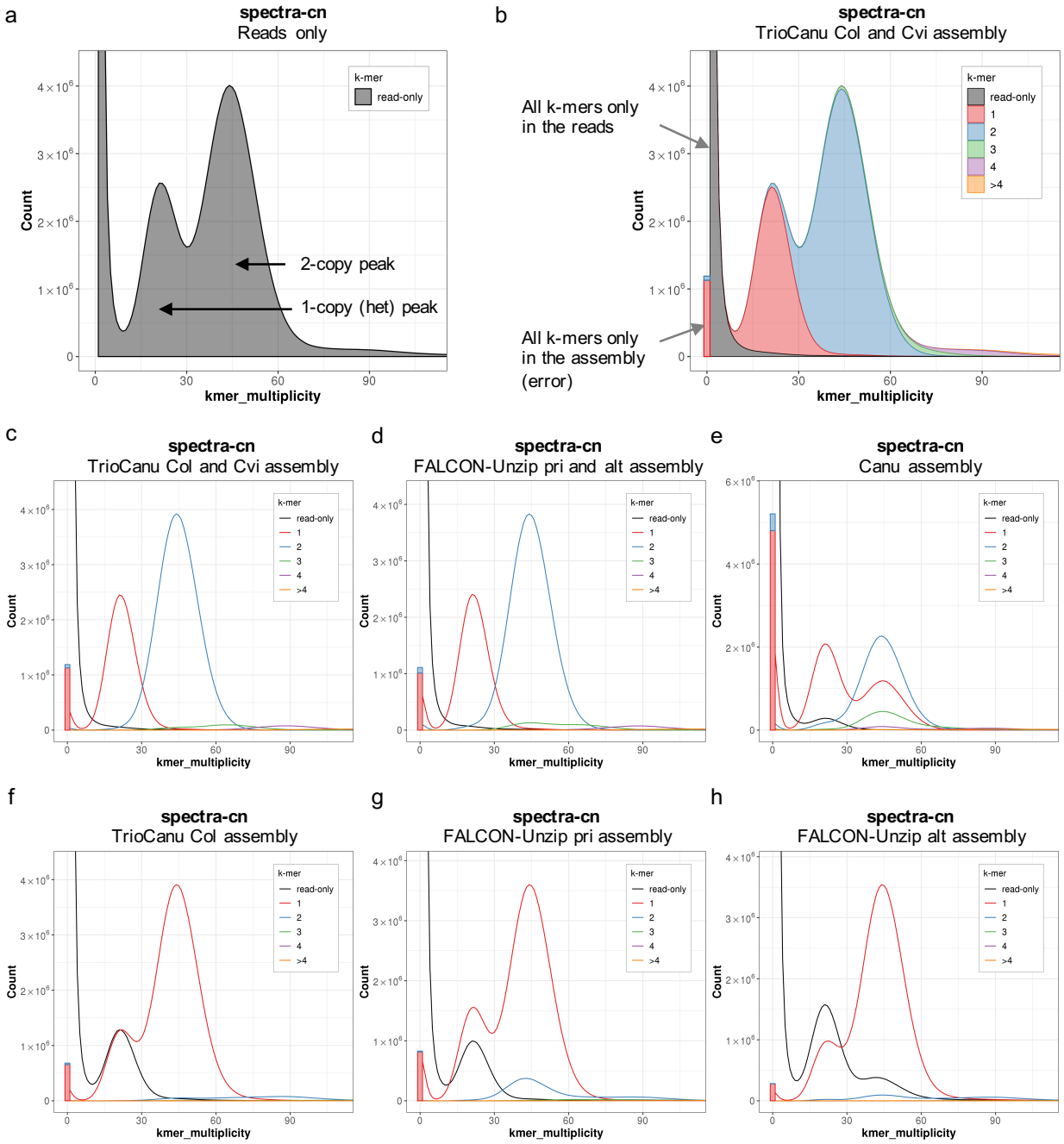


Fig. 1 | Merqury copy number spectrum plots for haploid and diploid assemblies of an *Arabidopsis thaliana* F1 hybrid genome. (a) Histogram of k-mer multiplicity collected from Illumina reads. By default, Merqury connects the midpoint of each histogram bin with a line, giving the illusion of a smooth curve. The first peak represents 1-copy (heterozygous) k-mers in the genome, and the second peak represents 2-copy k-mers originating from homozygous sequence or haplotype-specific duplications. Depth of sequencing coverage determines where these peaks appear. In this example, sequencing coverage is approximately 45x, corresponding to the 2-copy peak. (b) Copy number spectrum (spectra-cn) of the same k-mers in **a** plotted as stacked histograms colored by the copy numbers found in the combined TrioCanu assembly. The assembly k-mers absent from the read set (likely to be base errors in the assembly) are plotted as a bar at zero multiplicity, colored by the copy numbers found in the assembly. (c) Unstacked histograms of **b** for visualizing the distribution of k-mer counts per copy numbers found in the assembly. This plot shows an ideal pseudo haplotype assembly. (d) Spectra-cn plot of the combined FALCON-Unzip assembly. (e) Spectra-cn plot of the mixed-haplotype Canu assembly. Missing single copy k-mers (black) and k-mers from artificial duplications (green, purple, yellow in 30–60x) are noticeable. Note this assembly was not polished and so has a larger error k-mer bar. (f) Spectra-cn plot of the TrioCanu Col haplotype assembly. Half the single copy k-mers are missing and found in the other haplotype (black). Two-copy k-mers are found once (red) in each haplotype assembly. (g) Spectra-cn plot of the FALCON-Unzip primary assembly. (h) Spectra-cn plot of the FALCON-Unzip alternate assembly.

113 The bar at the origin of the plots represents k-mers found only in the assembly. From these
114 k-mers, we can estimate an assembly consensus quality value (QV), which represents a log-
115 scaled probability of error for the consensus base calls (**Methods**). Higher QVs indicate a more
116 accurate consensus, where Q30 corresponds to 99.9% accuracy, Q40 to 99.99%, etc. The trio-
117 binned assembly has QV scores of 35~36 for each haploid assembly, and 35.7 for the combined
118 version. The FALCON-Unzip assembly has a similar QV score of 35~38 for each haplotype, and
119 36 for the combined version. The error k-mer bar for Canu is much higher than the other two
120 diploid assemblies, as we omitted signal-level polishing[33] from this assembly to show an
121 intermediate assembly product. The estimated QV for this assembly is 29.3 (**Table 1**).

122 To better visualize the k-mer distribution by each copy number found in an assembly, we also
123 provide unstacked versions of the spectra-cn histograms (**Fig. 1c-e**). The 1-copy (heterozygous)
124 k-mers appear once in the combined TrioCanu assemblies (red), and 2-copy k-mers twice (blue)
125 as expected (**Fig. 1c**). The partially phased FALCON-Unzip assembly shows a similar distribution
126 to TrioCanu (**Fig. 1d**), indicative of good k-mer completeness. The lower fraction of 2-copy k-mers
127 found three times in the assembly (green hump under the 2-copy peak) indicates fewer false
128 duplications in the TrioCanu compared to FALCON-Unzip and Canu.

129 When generating the same spectrum on the mixed-haplotype Canu assembly (**Fig. 1e**), we
130 can see the assembly has only partially assembled both haplotypes and there is a higher fraction
131 of k-mers (37 kbp) missing from the assembly (black). In addition, the assembly has artificial
132 duplications inflating the assembly size. The plot also shows fewer 2-copy k-mers (blue peak)
133 compared to the other assemblies, and a significant fraction of 2-copy k-mers appearing only
134 once in the assembly (second red peak), suggesting the mixed-haplotype assembly has partially
135 collapsed the haplotypes. Based on the number of 2-copy k-mers found twice in the assembly,
136 we estimate 42.7 Mbp of homozygous sequence remains un-collapsed, typically at the boundary

137 of heterozygous and homozygous alleles. This partial separation is also evident in the higher
138 number of k-mers appearing three (green), four (purple), and more (yellow) times in the assembly,
139 representing 10.8 million artificially duplicated k-mers that would need to be identified and
140 removed to form a fully collapsed, haploid assembly. This example highlights the benefits of the
141 TrioCanu and FALCON-Unzip approaches for this heterozygous genome. Additional processing
142 with a tool such as purge_dups[34] would be required to convert Canu's mixed haplotypes into a
143 pseudo-haplotype.

144 The spectra-cn plots can also be useful for evaluating haploid assemblies, e.g. of a single
145 haplotype from a diploid genome. When plotting the same k-mer spectrum on one haplotype
146 ("Col" in this case), we can see both 1-copy and 2-copy k-mers are now observed just once in the
147 assembly (**Fig. 1f**, red histogram). This is because the read set represents the full diploid genome,
148 while the assembly isolates a single haplotype. The usual 1-copy peak is now exactly half the size
149 and perfectly overlaps with a peak of missing k-mers (black) that belong to the other haplotype
150 ("Cvi"). In comparison, the pseudo-haplotype FALCON-Unzip primary and alternate assemblies
151 (**Fig. 1g-h**) show imbalanced peaks, with more assembly k-mers appearing in both the 1- and 2-
152 copy peaks than expected. This suggests that the FALCON-Unzip is erroneously including
153 sequences from both haplotypes into the primary pseudo-haplotype. A similar portion of 1-copy
154 k-mers are missing from the alternate assembly (imbalanced red and black peaks), suggesting
155 that the alternate haplotype is missing some heterozygous variants.

156 ***Assembly spectrum***

157 In the above FALCON-Unzip pseudo-haplotype example (**Fig. 1g-h**), it is possible to infer that
158 the missing sequences in the alternate assembly are likely found in the primary assembly.
159 However, if there are shared sequences between the two assemblies, it is difficult to know the
160 exact sequence composition. To better address this question, we introduce a new method to show

161 the shared and specific k-mers in each assembly (spectra-asm), instead of showing the overall
162 copy-numbers (**Fig. 2**). This plot is helpful for measuring diploid assembly completeness as it
163 shows the fraction of k-mers specific to both the primary and alternate assemblies. For example,
164 a perfectly assembled diploid genome is expected to have a balance of k-mers specific to each
165 haplotype representing the heterozygous alleles (exceptions to this include sex chromosomes of
166 different size). The spectra-asm of the TrioCanu combined assembly (**Fig. 2a**) shows such an
167 example, where 1-copy k-mers are specific to each haplotype assembly (red and blue), and the
168 2-copy k-mers are shared by both assemblies (green). In comparison, the FALCON-Unzip
169 assembly is imbalanced, with more 1-copy and 2-copy k-mers in the primary assembly than
170 expected (**Fig. 2b**). This imbalance is also evident from Merqury's k-mer completeness metrics
171 for the primary and alternate assemblies (**Table 1**, completeness "all"). By stacking the
172 histograms, we can confirm the primary assembly contains all the missing k-mers from **Fig. 1h** in
173 the 2-copy peak. Compared to TrioCanu and FALCON-Unzip, Canu does not partition its output
174 into primary and alternates, and so k-mers from both haplotypes are present in the combined
175 assembly. However, the spectra-asm plot shows a few 1-copy k-mers are missing from the
176 assembly (**Fig. 2c**). This is in concordance with the lower k-mer completeness score in the Canu
177 assembly (95.7%) compared to the other diploid assemblies (98.3% and 97.8%, **Table 1**), and
178 indicates that some heterozygous variants are missing, perhaps as a result of the low-quality,
179 unpolished consensus.

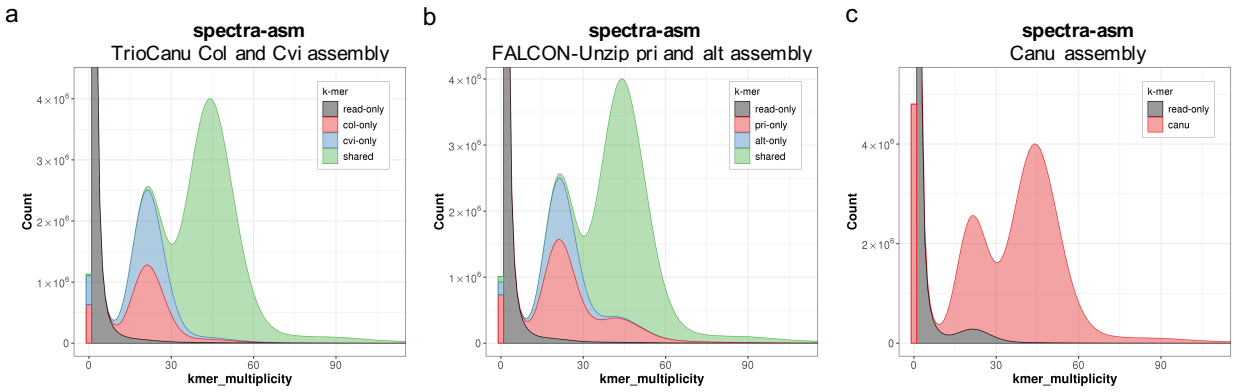


Fig. 2 | Merqury assembly spectrum plots for evaluating k-mer completeness. K-mers are colored by their presence in the reads and primary/alternate assemblies. (a) Distinct k-mer assembly spectrum (spectra-asm) plot of both TrioCanu Col and Cvi haplotype assemblies. This plot shows the assembly specific (red and blue) and shared portion of k-mers (green). (b) Spectra-asm plot of the FALCON-Unzip assembly. The primary assembly has more k-mers (red) compared to the alternate assembly (blue). (c) Spectra-asm plot of the Canu assembly which is a mixture of both haplotypes. A small fraction of 1-copy k-mers is found only in the reads (black peak around 12~30x), which represents heterozygous variants missing from the assembly.

Haplotype-specific k-mers (hap-mers) from trios

We define a "hap-mer" as a haplotype-specific k-mer that appears exclusively, one or more times, on a single haplotype of the genome. When parental genomes are available, we can use inheritance to estimate a set of hap-mers for the child and evaluate haplotype completeness of the assembly. Merqury identifies hap-mers as the set of inherited, parental-specific k-mers (**Fig. 3a**). Using the parental-specific markers alone may be sufficient for many cases; however, we have found it useful to specifically consider the inherited markers, as only half of the parental specific k-mers are inherited and the non-inherited markers may match an erroneous k-mer in the assembly by chance. Note, this is still only an estimate of the true set of hap-mers in the child's genome, which can also be affected by *de novo* variants or heterozygous variants that were differentially inherited from the parents. We have implemented efficient k-mer set operations (union, intersection, subtraction, etc.) within Meryl for computing hap-mers and other useful k-mer sets (**Fig. 3b, Methods**). For the *A. thaliana* F1 hybrid genome, we identified hap-mers directly from the genomes of the parental strains. K-mers were grouped based on their presence in the F1 reads alone, the maternal haplotype, the paternal haplotype, or both (**Fig. 3c**). Because of the

195 high heterozygosity of the F1 (estimated at 0.99% by GenomeScope[35]), many of the F1's k-mers
196 are hap-mers. A human genome, in comparison, has many fewer hap-mers, with most of the k-
197 mers shared between both haplotypes (**Fig. 3d**).

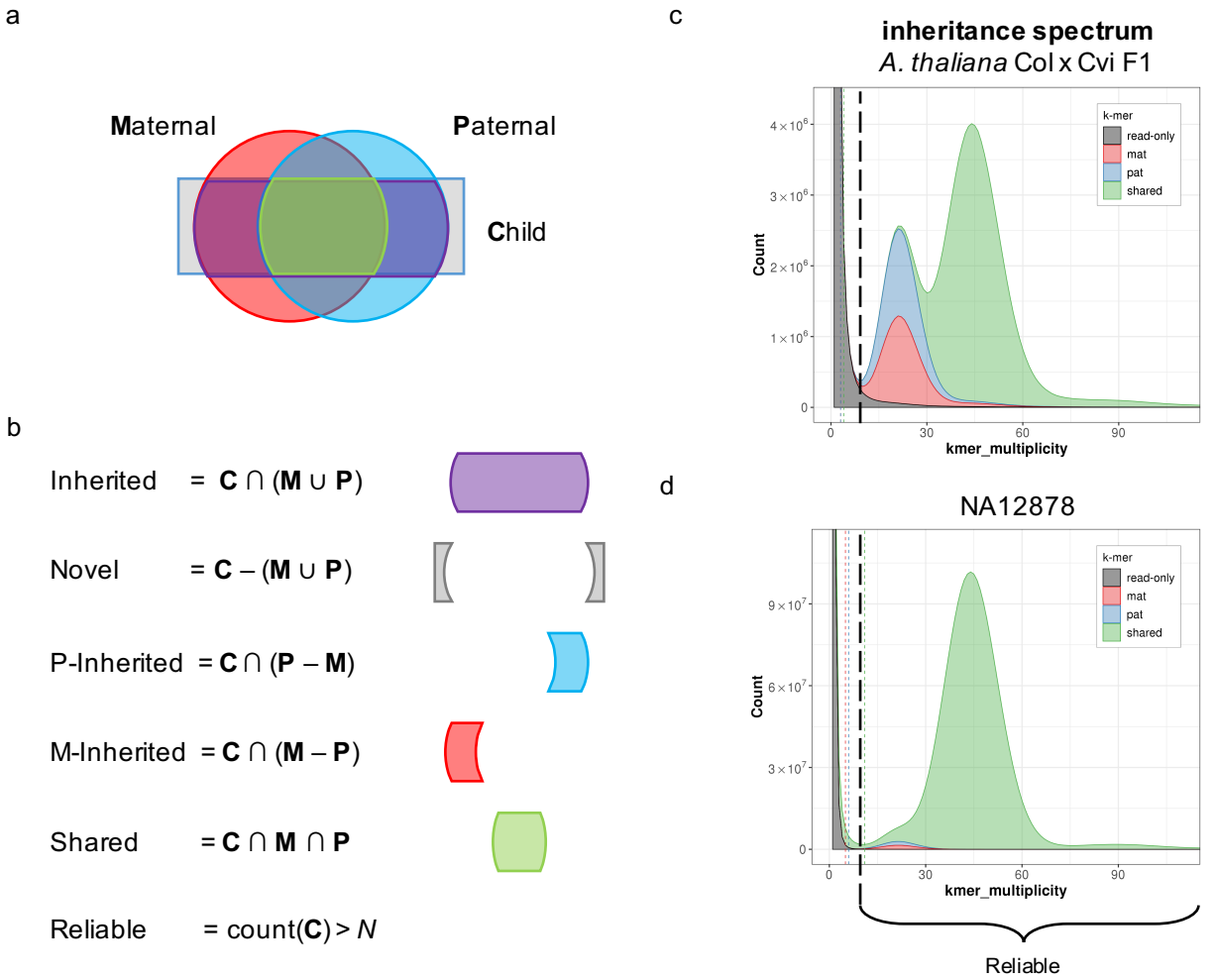
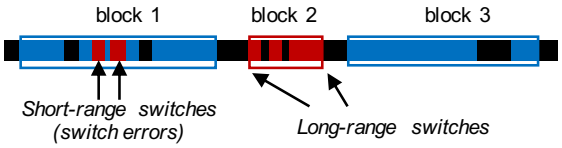


Fig. 3 | Merqury set operations for generating haplotype specific hap-mers and reliable k-mers. (a) Venn diagram of maternal, paternal, and child k-mer sets. Inherited haplotype-specific k-mers (hap-mers) are estimated from the parental and child k-mer sets. Roughly half of the parental-specific k-mers are inherited by a child. (b) Set operation examples used in Meryl to compute hap-mers and other k-mer sets. (c-d) Stacked k-mer multiplicity of the child's read set, colored by inheritance. K-mers are colored by maternal (red) paternal (blue) and shared between parents (green). K-mers only seen in the child's reads (black) are mostly from low-copy sequencing errors or k-mers arising from *de novo* variants in the child. Reliable k-mer thresholds used for generating completeness scores for all k-mers (black) and hap-mers (red and blue) are marked by dashed lines.

a



b

Phase Blocks	Total	# Blocks	Max	NG50	Switch (%)
TrioCanu (Col)	123.2	412	12.6	3.6	0.34
TrioCanu (Cvi)	122.0	340	11.0	5.5	0.28
FALCON-Unzip (pri)	139.2	343.0	10.2	3.1	0.25
FALCON-Unzip (alt)	104.7	393.0	8.5	2.5	0.32
Canu	243.2	5,612	1.0	0.1	3.45

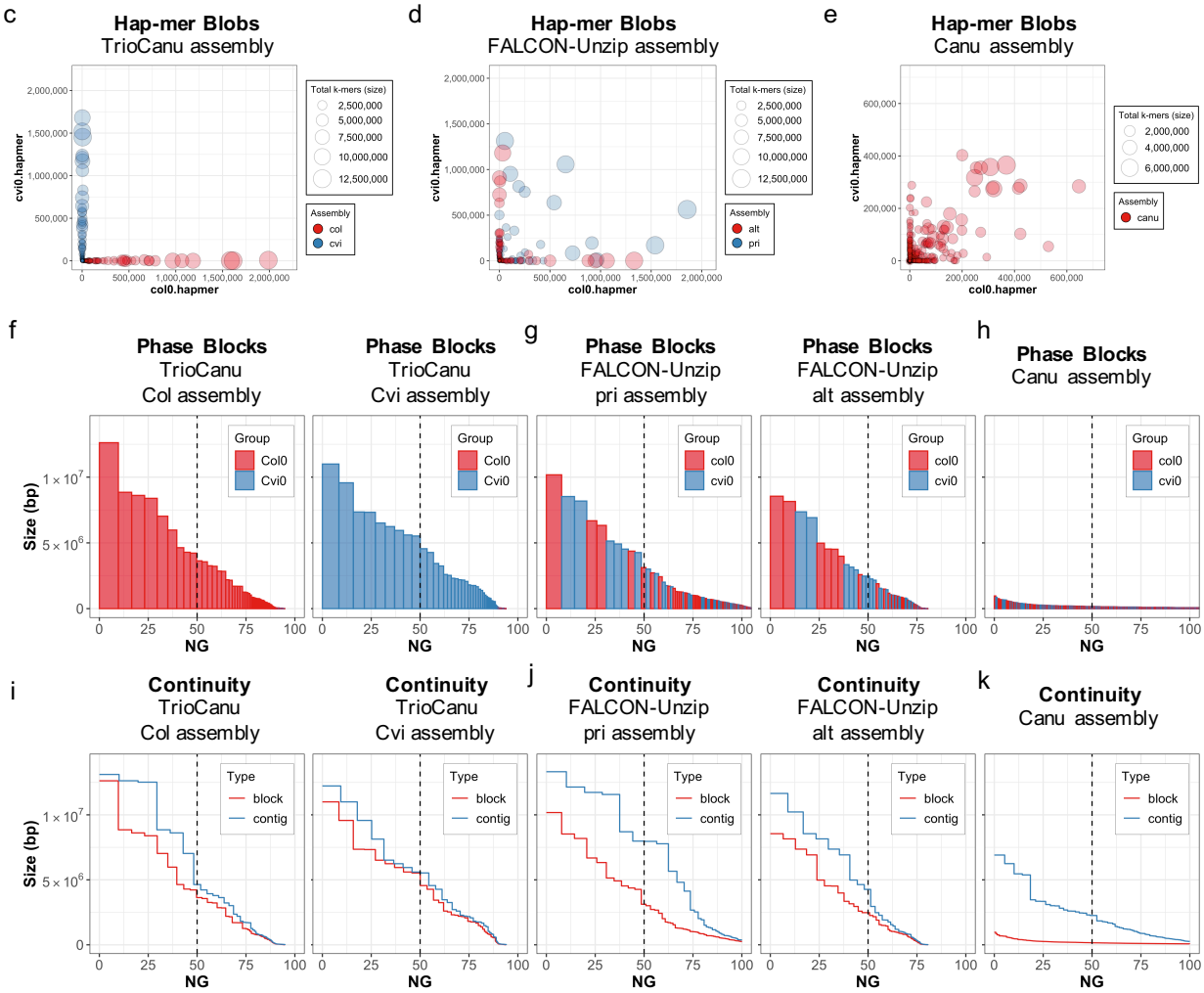


Fig. 4 | Merqury hap-mer plots for evaluating haplotype phasing. (a) Example of phase blocks and switches. Blue and red bars are paternal or maternal hap-mers found in the assembly. A phase block is defined by at least two hap-mers (markers) from the same haplotype. Short-range switches are allowed in between markers, in defined ranges. Two consecutive red markers within a certain range are marked as short-range switches and counted for switch errors in block 1. As the red markers are consecutively found over a certain range, or in numbers above a certain threshold, a separate block is formed. Each switch between blocks is counted as a long-range switch. (b) Phase block statistics of the haploid assemblies with switch errors, allowing at most 100 switches within 20 kbp. (c) Hap-mer blob plot of the TrioCanu assembly. Red blobs represent Col haplotype contigs, while blue blobs are the Cvi haplotype. Blob size is proportional to contig size, and each blob/contig is plotted according to the number of contained Col (x values) and Cvi (y value) hap-mers. Col-specific k-mers are found in the Col assembly with almost no Cvi-specific k-mers, while Cvi k-mers are found in the Cvi assembly with almost no Col k-mers. (d-e) Blob plots for FALCON-Unzip and Canu assemblies show that most contigs are a mix of sequences from both haplotypes, but FALCON-Unzip preserves phase within its alternate contigs, as designed. (f) Phase block NG* plots of the haplotype resolved Col (left) and Cvi (right) assembly, sorted by size. X-axis is the percentage of the genome size (*) covered by phase blocks of this size or larger (Y-axis). Blocks from the wrong haplotype are very small and almost entirely absent. (g-h) Phase block NG* plot of the (g) FALCON-Unzip and (h) Canu assemblies. Col and Cvi phase blocks are distributed evenly, as is typical for pseudo-haplotype assemblies. (i-k) Phase block and contig NG* plots show the relative continuity of (i) TrioCanu, (j) FALCON-Unzip, and (k) Canu assemblies. Phase block sizes are similar to the contig sizes in **i**. Phase blocks are much shorter than the contigs in **k**, because of the frequent haplotype switches in the contigs.

198 ***Evaluating phasing completeness with hap-mers***

199 Hap-mers are used to determine phase blocks in Merqury, where a block is defined to be a
200 consistent set of markers originating from the same haplotype. To account for minor base-level
201 errors in the assembly, we do allow some short-range switches to occur within a block, so long
202 as the phase switches back shortly thereafter (**Fig. 4a**). A benefit of this k-mer approach is that
203 Merqury does not need to rely on the phase blocks as identified by the assembler or a reference
204 variant callset, and can quickly compute the blocks on each assembly directly using only the
205 observed haplotype markers. Applying this method to the TrioCanu assembly reported NG50
206 phase block sizes of 3.6 Mbp and 5.5 Mbp with 0.3% per-block switch rate when allowing at most
207 100 consecutive switches within 20 kbp (**Fig. 4b**). The FALCON-Unzip assembly has slightly
208 shorter phase block sizes of 3.1 Mbp and 2.5 Mbp with a similar switch error rate of 0.3%. The
209 Canu assembly had more frequent long-range switches among haplotypes, resulting in NG50
210 phase blocks of 100 kbp.

211 Visualizing hap-mer presence in each haplotype assembly is also useful to detect overall
212 phase consistency. When counting Col- and Cvi-specific k-mers in contigs of the TrioCanu
213 assembly (**Fig. 4c**), each contig was successfully separated by haplotype as expected. That is,
214 the Col markers were observed in the Col haplotype assembly, with almost no contaminating Cvi
215 markers, and vise versa. As such, a haplotype resolved assembly with almost no haplotype
216 switches is expected to have a similar plot with each blob (contig) close to the corresponding plot
217 axis. The FALCON-Unzip alternate contigs maintain haplotype consistency (**Fig. 4d**), but the
218 primary pseudo-haplotype contigs are a mixture of both haplotypes. The Canu assembly appears
219 to mix haplotypes in all but the smallest contigs (**Fig. 4e**), and does not partition the resulting
220 assembly into primary and alternate contig sets with post-processing with Purge_dups.

When plotting phase blocks sorted by size, the blocks originating from the wrong haplotype were very small and almost negligible in the TrioCanu assembly (**Fig. 4f**). In contrast, the phase blocks were highly mixed in the pseudo-haplotype assemblies, with the larger contigs being more likely to contain markers from both haplotypes (**Fig. 4g-h**). Plotting the contig and block sizes together shows that the trio-binned phase blocks are very similar in size to the trio-binned contigs (**Fig. 4i**). In comparison, the phase blocks were shorter than the contigs in the FALCON-Unzip assemblies (**Fig. 4j**), showing relatively good phasing performance. The phase blocks were much shorter in the Canu contigs, indicating frequent block switches between haplotypes (**Fig. 4k**) since Canu does not attempt to preserve long-range phasing.

Another useful feature in Merqury is that all hap-mers, erroneous k-mers, and phase blocks can be visualized as genome tracks along the assembly. **Fig. 5** shows an example of a 60 kbp region in the mixed-haplotype Canu assembly, where haplotypes are observed switching from Cvi (blue) to Col (red), resulting in numerous base errors (grey). This illustrates how failure to separate haplotypes can result in increased base error, as the two haplotypes are improperly combined into a single consensus sequence resulting in artifactual variants. Two phase blocks (Cvi and Col) were found in this region using the default threshold of 100 consecutive k-mer switches allowed per every 20 kbp. However, phase blocks can be more stringently measured by defining the short-range switch allowance (e.g. 10 per 20 kbp, **Fig. 5**, bottom track), resulting in lower NG50 phase block size (100 kbp decreases to 33 kbp). In contrast, the per-block switch error rate decreased from 3.4% to 0.47%, making each block a more reliable haplotype. Note the per-block switch error rate is defined as the fraction of k-mer markers within a block that are assigned to the wrong haplotype, thus accounting for all short-range marker switches in a block.

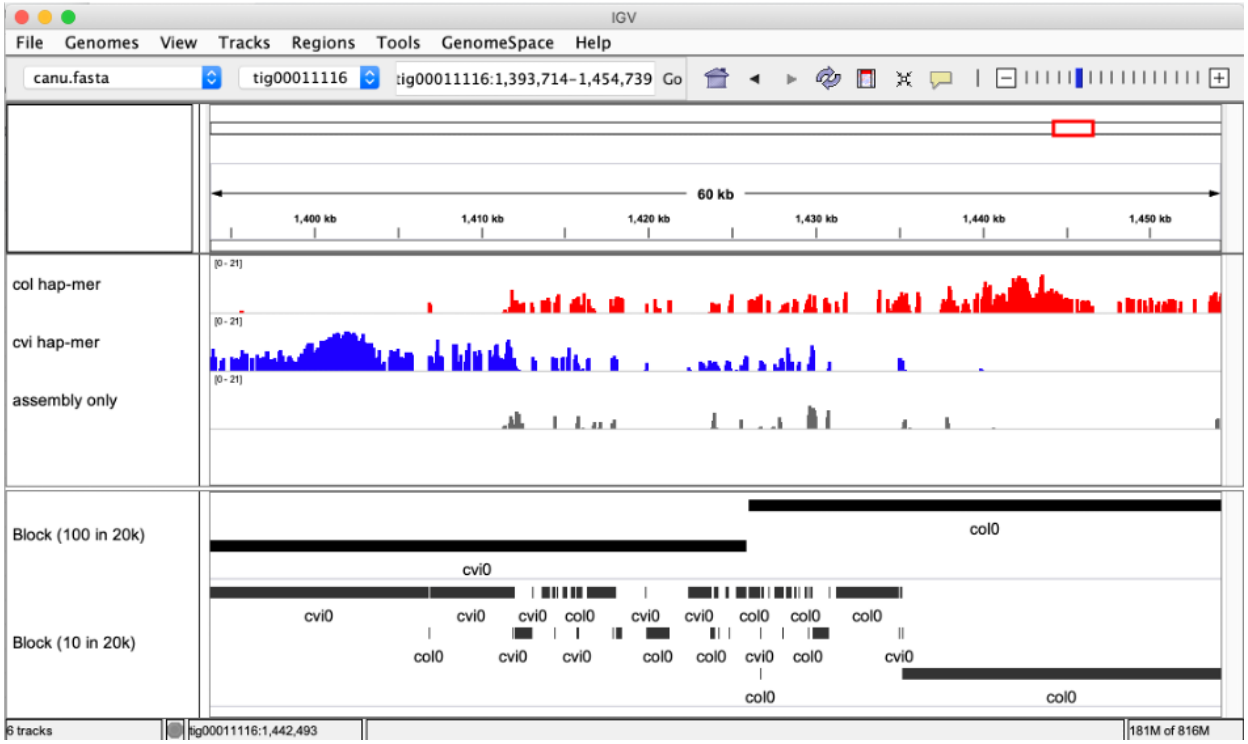


Fig. 5 | Example k-mer and phasing tracks provided by Merqury. Hap-mer density is provided in tiled data format (.tdf) browsable with the Integrated Genomics Viewer (IGV)[36,37]. This figure shows a region where haplotype blocks are switching within one of the Canu contigs. Hap-mer tracks show haplotype switches from Cvi (blue) to Col (red), along with k-mers found only in the assembly (grey), which are likely caused by erroneous consensus bases. Phase blocks (black) are shown for both relaxed (100 consecutive switches allowed per 20 kbp) and strict (10 per 20 kbp) switching thresholds.

243 ***Benchmarking on a human genome (NA12878)***

244 To benchmark Merqury on a large genome, we applied it to the TrioCanu human (NA12878)
245 assembly from Koren *et al.*[10] and estimated the consensus quality as Q31 for each haplotype
246 (**Table 1**). The alternative variant calling approach reported 2.1 million bases of errors, resulting
247 in a QV estimate of 34.1. However, the mapping-based approach excluded 212.2 Mbp of
248 assembled sequence because of too few (<3x) or too many (>600x) aligned reads. We argue that
249 Merqury's k-mer based approach is both more efficient and more accurate for evaluating
250 consensus accuracy.

251 Merqury required only 14.9 CPU h (9.1 h for k-mer counting, 4.7 h for merging, 1.1 h for
252 statistics) to evaluate QVs for both haplotypes. In contrast, the mapping-based QV estimate took
253 338.3 CPU h (2.2 h for indexing, 308 h for mapping, 12.6 h for merging, 12.8 h for variant calling,
254 and 2.5 h for coverage calculation and QV estimates). By excluding low and high-coverage
255 regions of the assembly, the mapping-based approach ignores regions of the assembly likely to
256 be enriched for error. For example, low coverage regions can be caused by regions of high error
257 rates that makes it difficult to map short reads. High coverage regions are typically caused by
258 repeats that can be collapsed, and therefore incorrect, in the assembly. Thus, a substantial
259 number of errors may be excluded from the accuracy statistics if one considers only the mappable
260 portion of the assembly. This matches with our observation that the mapping-based estimates
261 always overestimate QV compared to the k-mer based approach (**Table 1**).

262 Exact k-mer counting is currently the most resource intensive step of Merqury, requiring a
263 maximum 21 GB of memory using 25.5 GB of disk space on NA12878 (**Table 2**). While this step
264 can be parallelized across multiple nodes and cores, QV statistics can be also estimated from
265 subsampled k-mers with lower memory and disk requirements using Mash Screen[38]. Because
266 Merqury's QV estimation is based on Mash's k-mer containment score (**Methods**), the Meryl and

267 Mash counting methods are interchangeable. In comparison to Meryl, Mash streams sequencing
268 reads from disk and compares them against only a small subset of k-mers in the assembly. This
269 avoids the need for a large table of k-mers, but at the same time ignores copy number information.
270 As a result, we observed that Mash QV estimates were slightly lower than exact counting (Meryl)
271 for each haplotype, and even lower when both haplotypes were combined (**Table 1**). This is
272 because the shared k-mers between the two haplotype assemblies are considered only once by
273 Mash, resulting in an underestimate of the QV score (e.g. if a 2-copy k-mer appears in just one
274 haplotype, it is considered "correct" by Mash). The Mash approach also cannot investigate
275 positional base errors (**Figure 5**) and many of the other analyses presented here, but is provided
276 as an alternative to Meryl for QV estimation in cases where disk and memory resources are
277 limited.

Table 2. Merqury runtime, memory, and disk requirements for a human genome.

	Count	Union-sum	QV (Meryl)	QV (Mash)
cpus x nodes	32 x 24	48 x 1	24 x 1	24 x 1
wall clock time	6m 52s / node	7m 43s	14m 13s	3h 36m 17s
cpu time	9.1 h	4.7 h	1.1 h	19.0 h
memory	21.2 G	7.0 G	10.56 G	2.6 G
storage	90 G (fastq.gz)	N/A	48 G	90 G (fastq.gz)
intermediates	1.8 G x 24	48 G	25.5 G	23.1 M

All statistics are for the diploid human genome NA12878. QV Meryl reports the Merqury QV estimates generated on the full k-mer databases, whereas QV Mash reports the Merqury QV estimates generated only from the subsampled Mash sketches. Runtimes were measured on Intel(R) Xeon(R) Gold 6140 CPU, with 2.30GHz. Storage requirements represent gzipped FASTQ files for counting and QV (Mash), and a binary database for QV (Meryl). h, hours; m, minutes; s, seconds; G, Giga bytes.

278 Next, we evaluated k-mers for NA12878. The hap-mer spectrum of NA12878 revealed a
279 higher fraction of shared k-mers in the 1-copy peak (**Fig. 3d**), indicating that some heterozygous
280 variants in the child are shared by both parents. We do not see such a strong effect in *A. thaliana*,
281 because the parents were heavily inbred and contained few heterozygous variants of their own.
282 In contrast, the *A. thaliana* F1 hybrid was deliberately outbred, which is evident by the dramatically
283 taller 1-copy peak versus NA12878 (0.99% vs. 0.12% heterozygosity).

284 Haplotype-specific k-mers are convenient to obtain haplotype precision (PPV) and recall
285 (completeness) statistics based on how many of the expected parental k-mers are observed in
286 the child's haplotype-resolved diploid assembly (**Fig. 6**). To demonstrate, we built genomic k-mer
287 databases for NA12878 and her parents, totaling 18.4 and 19.9 million inherited hap-mers for the
288 paternal and maternal haplotypes, respectively. When comparing to the haplotype-resolved
289 assemblies, the maternal haplotype assembly recovered 90.7% of the maternal hap-mers (**Table**
290 **1** and **Fig. 6a**), and the paternal assembly recovered 91.0% of the paternal hap-mers (**Table 1**
291 and **Fig. 6b**). Likewise, by considering the other haplotype's markers as false positives (i.e.
292 paternal hap-mers found in the maternal assembly), the precision of the maternal and paternal
293 assemblies was 99.1% and 98.8%, respectively, with only 160~200 markers appearing in the
294 incorrect haplotype (**Table 1**, PPV values). This evaluation excludes all k-mers found only in the
295 assembly (errors), which if considered false positives, would further lower the precision.

296 To compare Merqury's results with an alternative approach, we considered an Illumina-based
297 platinum callset for NA12878[26] that includes 3.4 million heterozygous SNPs in regions where both
298 haplotype assemblies align to the reference (hg38). Calling SNPs directly from the haplotype
299 assemblies against hg38 recovered 3.2 million variants, or 93.5% sensitivity, which is slightly
300 higher than the Merqury based estimate of ~91%. This discrepancy is likely due to Merqury's
301 ability to measure complex regions of the assembly not easily measurable by a mapping-based
302 analysis. In contrast, the SNP-based measurement of precision was only 86.1%, likely due to the

303 low base accuracy (Q31) of the assemblies, where consensus errors are called as false-positive

304 heterozygous SNPs. Thus, it is important to consider both the k-mer QV and precision estimates

305 when evaluating the accuracy of a diploid assembly.

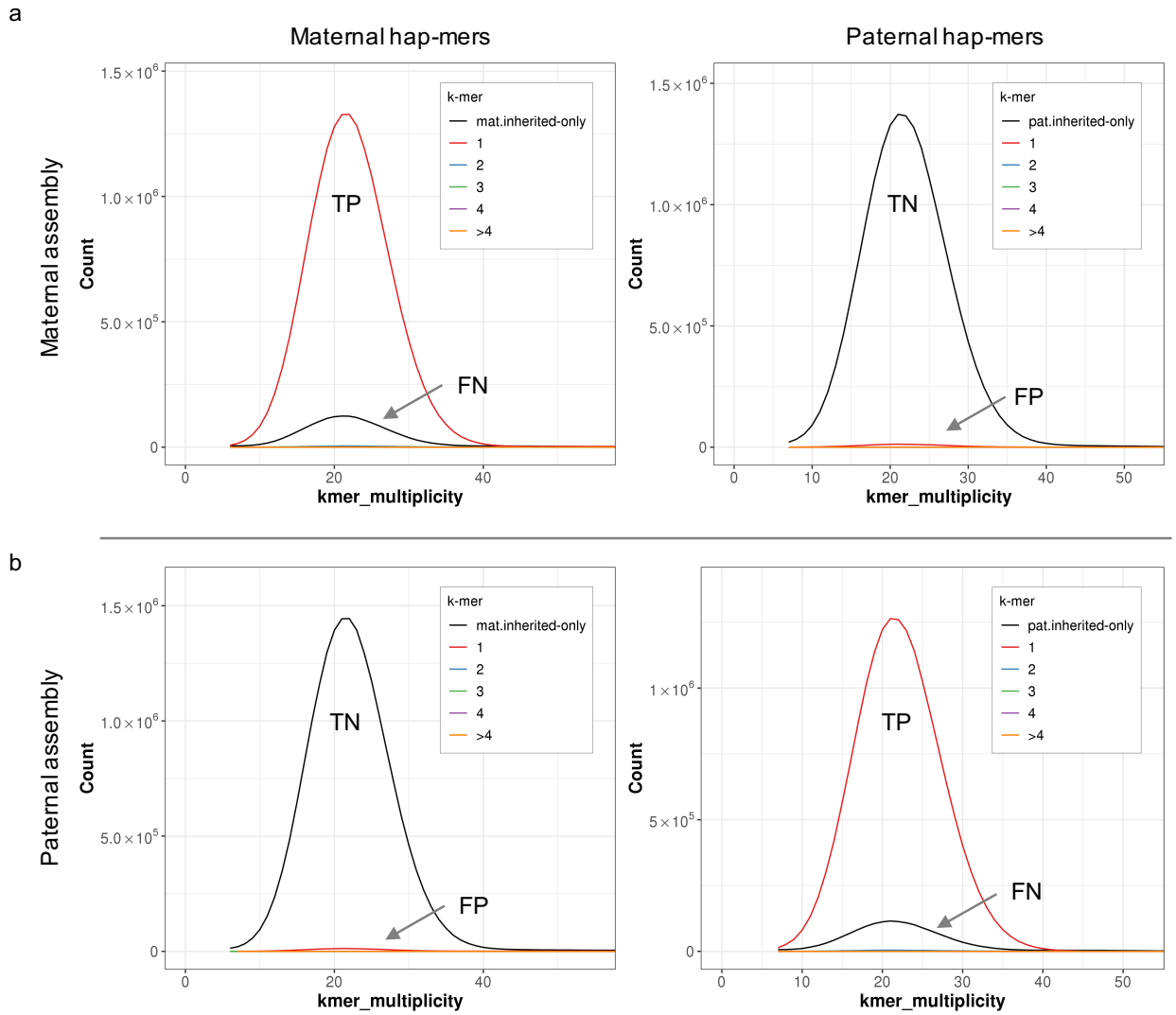


Fig. 6 | Evaluating haplotype completeness for diploid assemblies using Merqury. Spectra-cn plots of the maternal and paternal assemblies of the TrioCanu NA12878 assembly are displayed as a confusion matrix. (a) Hap-mer spectra-cn plot of the maternal assembly. All maternal hap-mers are expected to be found in the maternal (top left), with no paternal hap-mers (top right). Any maternal hap-mers not found in this assembly are missing hap-mers (top left, black). Paternal hap-mers (top right, red) are false positive phasing errors. (b) Hap-mer spectra-cn plot of the paternal assembly. This time, no maternal hap-mers should be seen (bottom left), and only paternal markers are expected to be present (bottom right). TP, true positives; FN, false negatives; TN, true negatives; FP, false positives.

306 **Conclusion**

307 We have developed Merqury, a reference-free assembly evaluation toolkit based on efficient
308 k-mer based methods. Merqury builds upon the k-mer spectra ideas of Mapleson *et al*[17], and
309 introduces novel methods and plots for measuring assembly quality (QV), completeness, and
310 phasing. Using k-mer count spectra, Merqury can reveal copy-number errors in an assembly and
311 accurately measure both assembly completeness and consensus quality. When parental k-mers
312 are available, Merqury can also measure phasing accuracy and haplotype completeness. In
313 addition to validation statistics, Merqury provides a number of graphs for interpreting assembly
314 quality. In the process of developing Merqury, we have also extended Meryl with generalized k-
315 mer counting, querying, and set operations that will be useful for other k-mer based analyses (**Fig.**
316 **3b**).

317 Merqury is able to evaluate assemblies from any sequencing technology, and works best
318 when high-accuracy sequencing reads are available from the assembled individual and its
319 parents. This read set serves as independent validation of the assembly, which is typically based
320 on less-accurate, long-read sequencing. If high-accuracy reads are not available from the
321 assembled individual, read sets from the parents can be used as a replacement for measuring
322 quality values. This assumes all k-mers in the child are found in the parental genomes, ignoring
323 the small fraction of k-mers from *de novo* mutations in the child. Although we currently recommend
324 using Illumina data for the k-mer based validation, it may require special library preparation to
325 minimize sequencing biases[39-41]. We note that Merqury's methods are general and would be
326 compatible with any reasonable high-accuracy, high-throughput sequencing technology.

327 Hap-mers are currently computed by a simple set operation, similar to trio-binning[10]. A higher
328 portion of hap-mers are identifiable when the parents are divergent, with minimal shared ancestry.
329 Our hap-mer spectrum plots (**Fig. 3c-d**) show the 1-copy haplotype markers that are specific to

330 each parent, which may not be as prevalent for genomes of low heterozygosity, such as humans.
331 Here we have used trios of diploid genomes to identify haplotype-specific markers, but Merqury's
332 methods are extensible to polyploid genomes and it is possible to identify hap-mers using
333 orthogonal datasets (e.g. Hi-C, Strand-seq). To support alternative k-mer classification methods,
334 Merqury is designed to receive any pre-computed hap-mer set as input.

335 We argue that Merqury's k-mer-based method provides better haplotype completeness
336 estimates, because it does not rely on a reference genome. Mapping to a reference can be biased
337 by mis-mapping to repetitive or low-quality regions of the assembly. Moreover, k-mers naturally
338 capture heterozygous insertion and deletion variants and are thus immune to the problems of
339 calling these types of variants with a reference mapping approach. For example, consortiums
340 such as the GA4GH exclude all variant calls within complex, repetitive regions of the genome[25].
341 In contrast, k-mers inherently capture genetic context, regardless of the structural complexity
342 surrounding them in the genome. Moreover, k-mers are especially robust for evaluating
343 sequences in highly diverged sequences, where mapping based approaches cannot map reads
344 to call variants.

345 Lastly, Merqury provides an efficient way of determining phase blocks in diploid assemblies.
346 In the past, phase blocks were defined based on heterozygous SNPs, measured by aligning each
347 haplotype to one another[9] or by mapping to a reference genome[8]. These alignment-based
348 approaches may not consider the full genome when the identity between the two haplotypes is
349 lower than the alignment threshold, or the alignment is confused by genomic repeats. Moreover,
350 a reference genome may not represent the entire haplotype of an assembled individual, thereby
351 omitting haplotype-specific sequences from the analysis. The phase blocks measured by Merqury
352 are generated regardless of the haplotype being assembled, and provide more reliable phasing
353 information for allele-specific studies.

354 ## Methods

355 ### Counting k-mers with Meryl

356 Meryl is a tool for counting and working with sets of k-mers that was originally developed for
357 use in the Celera Assembler[42] and has since been adopted and maintained as part of Canu[32].
358 Here we have improved Meryl's efficiency and extended it to support a variety of functions useful
359 for k-mer-based assembly validation. A set of k-mers and their associated counts is termed a k-
360 mer database. The count is the number of times a k-mer occurs in some collection of sequences.
361 The k-mer database is stored in sorted order, similar to words in a dictionary. Meryl comprises
362 three modules: one for generating k-mer databases, one for filtering and combining databases,
363 and one for searching databases.

364 The counting module uses two different algorithms: one for k-mers up to size 16 and the other
365 for k-mers up to size 64. For small k-mers, Meryl directly counts the number of times each k-mer
366 occurs in the input sequences. An array of 4^k 16-bit integers is allocated. Each k-mer is converted
367 to an integer index into the array, and the cell for that k-mer is incremented. When any cell
368 exceeds the maximum possible value that can be represented, the width of the array is extended
369 by allocating a supplementary array of 4^k bits. For large k-mers, Meryl generates lists of all the k-
370 mers present in the input sequences, sorts each list, then scans each to determine how many
371 times a specific k-mer occurs. Each k-mer is split into a prefix and a suffix. The prefix is used to
372 select a list, and only the suffix is added. A trade-off is made between a small prefix (which would
373 result in a only a few lists, each storing many suffixes) and a large prefix (which would result in
374 many lists). As we do not know how big each list will be, the lists must be able to grow as needed.
375 Each list is therefore an array of memory blocks where each block can store a few thousand k-
376 mers. While counting, the memory usage of the lists is tracked, and if a user-supplied memory
377 limit is reached, the lists are sorted, k-mers are counted, and output written to an intermediate

database. After all k-mers are processed, the intermediate, sorted databases are efficiently merged into one.

With one or more databases on disk, Meryl can filter or combine k-mers to create new databases. Each database is stored in 64 independent pieces, and each piece can be processed in parallel. Meryl can filter a database by count (e.g. less-than, greater-than or equal-to some user supplied constant), or by fraction of distinct k-mers in a database (e.g. the most common 5% of the k-mers). It can modify the count of every k-mer in a database by a constant (e.g., add 1, subtract 1, multiply by 2). Meryl can also output the union or intersection of multiple databases, setting the count of a k-mer to the minimum, maximum, sum of all copies of the k-mer or as the count of the first database. It can output the difference of databases (e.g. the k-mer occurs only in the first database) or the symmetric-difference (e.g. the k-mer occurs in exactly one database). Conveniently, any number of these operations can be combined into one command, using a reverse-polish-notation inspired format. The following example invocations are used in Merqury:

1. Write the k-mers that occur in both `db1.meryl` and `db2.meryl` to `db3.meryl`. Set the count of each output k-mer to the sum of the counts in the input k-mers: `meryl union-sum db1.meryl db2.meryl output db3.meryl`

2. Output k-mers that occur in both `db4.meryl` and `db5.meryl`, additionally requiring the k-mer in `db5.meryl` to be unique. The count of the output k-kmer is set to the count of the k-mer in the first input to the `intersect` operation, namely `db4.meryl`:

```
meryl    output    db6.meryl    \
  intersect                     \
    db4.meryl                   \
    [ equal-to 1 db5.meryl ]
```

3. For each k-mer in `asm.fasta`, output the (0-based) coordinate of the kmer in the sequence, the forward and reverse k-mer sequences, and the count of the k-mer in `db7.meryl`:

```
meryl-lookup -dump -sequence asm.fasta -mers db7.meryl
```

Meryl includes a C++ API to extend its functionality. For example, random lookups can be added using either the simple existence of a k-mer in a database, or the count associated with a k-mer. On the command line, lookups can return the number of k-mers a sequence shares with a database, a list of each k-mer in a sequence annotated with the count the k-mer has in a database, or a filtered list of input sequences based on the presence or absence of k-mers in the database.

Evaluating assemblies with Merqury

Copy number spectrum (spectra-cn plot)

Given a genome size G and tolerable collision rate p, an appropriate k can be computed as $k = \log_4 (G(1-p)/p)$[43]. However, alterative values of k may be desirable for different scenarios, e.g. increasing k has the effect of increasing the relative fraction of 1-copy k-mers, which may be useful for genomes with low heterozygosity, but this also increases the fraction of erroneous k-mers. Decreasing k reduces the fraction of erroneous k-mers, but increases the fraction of repetitive k-mers. Once an appropriate size if k is determined (typically 18~21), we count the canonical k-mers observed in the assembly and in the accurate, whole-genome read set. A typical k-mer spectrum for a heterozygous diploid genome consists of two primary peaks, one representing k-mers that are 1-copy in the diploid genome (heterozygous, on a single haplotype) and one representing those that are 2-copy in the diploid genome (homozygous, on both haplotypes or two copies on one haplotype) (**Fig. 1a**). The 2-copy k-mers appear with a frequency approximately equal to the average depth of sequencing coverage, where the 1-copy k-mers

424 appear with frequency approximately equal to half the sequencing coverage. If a genome is
425 entirely homozygous, only the 2-copy peak may appear, and if the genome is extremely
426 heterozygous, only the 1-copy peak may appear. With sufficient sequencing coverage (to
427 separate the peaks along the axis), and a proper choice of *k*, both peaks are visible for most
428 genomes. Using the multiplicity of the k-mer counts, and modeling the k-mer survival rate (i.e.
429 how many k-mers are unaffected by sequencing error), it is possible to predict the size and repeat
430 content of a genome from the k-mer spectrum alone[35].

431 The spectra-cn plot was introduced by Mapleson *et al.*[17], which colors k-mers of the read set
432 by their copy numbers in the assembly. In addition to the original stacked version of the spectra-
433 cn plot (**Fig. 1b**), we provide additional options to plot the unstacked copy number spectrum (**Fig.**
434 **1c**). We have found this style more useful for visually detecting abnormal k-mer copy numbers
435 and their distribution in an assembly.

436 *Assembly spectrum (spectra-asm plot)*

437 Similar to the spectra-cn analysis, we can color each k-mer in the read set by the assembly
438 in which it is found. This becomes useful when two haploid assemblies are evaluated. This way,
439 we can detect k-mers that are present only in one assembly, k-mers shared in both assemblies,
440 and k-mers not present in the assembly and only found in the read set (**Fig. 2**).

441 *K-mer completeness*

442 We define a "reliable k-mer" as a k-mer that is truly in the genome and unlikely to be caused
443 by sequencing error. With exact k-mer counts, it is easy to filter out low-copy k-mers that are likely
444 to represent sequencing errors. We use the same strategy as Koren *et al.*[10] to find the cutoff. In
445 brief, we take the histogram of the k-mer counts and set the multiplicity (number of times we see
446 the k-mer in the read set) as *x* and counts (number of k-mers with *x* multiplicity) as *y*. When

447 differentiating the histogram, we compute the slopes and the first k-mer multiplicity with a positive
448 slope defines the reliable k-mer threshold. Examples of these cutoffs are shown as dashed lines
449 in **Fig. 3c-d**. The k-mer completeness is calculated as the fraction of reliable k-mers in the read
450 set that are also found in the assembly. For repetitive genomes, erroneous read set k-mers can
451 sometimes appear above this threshold due to recurring errors in high-copy repeat families, but
452 this is a rare.

453 *Consensus quality (QV) estimation*

454 We can also use k-mers to estimate the frequency of consensus errors in the assembly. We
455 use a binomial model of k-mer survival, and assume all k-mers in the assembly should be found
456 at least once in the read set. Here, we use the containment score from Mash Screen[38] to estimate
457 consensus accuracy. In brief, we estimate the probability P that a base in the assembly is correct
458 as:

459 $$P = (K_{shared} / K_{total})^{1/k}$$

460 Where the K_{total} is the total number of k-mers found in an assembly and K_{shared} are the number of
461 shared k-mers between the assembly and the read set. If the read set is assumed to completely
462 cover the genome, any k-mer found only in the assembly ($K_{asm} = K_{total} - K_{shared}$) likely reflects a
463 base error in the assembly consensus. Hence, the error rate E can be defined as:

464 $$E = 1 - P = 1 - (1 - K_{asm} / K_{total})^{1/k}$$

465 Using this formula, the widely used Phred[44] quality score (often denoted as QV) can be computed
466 by treating the E as base error probability:

467 $$QV = -10 \log_{10} E$$

468 Note that this QV estimate depends on the coverage and quality of the read set. The assembly
469 QV will be underestimated if the read set does not contain all true k-mers of the genome due to
470 low or heavily biased sequencing coverage. Conversely, the QV may be overestimated if the read
471 set contains a high proportion of false k-mers (e.g. due to the combination of extreme coverage
472 and sequencing error). Lastly, the k-mer survival model assumes both k-mers and errors are
473 independent, but k-mers overlap one another and errors tend to cluster together. As a result,
474 choice of k also affects the QV estimates, with $k \approx 21$ recommended by Mash based on empirical
475 testing.

476 *Positions of k-mers for mis-assembly detection*

477 Merqury reports the positions of the k-mers found only in an assembly for further investigation
478 in .bed and .tdf formats that can be loaded into most genome browsers. In addition, the k-mers
479 found in unexpected copy numbers (i.e. false duplications) are also provided as .bed and .tdf files.
480 The format details are described at:

481 https://genome.ucsc.edu/FAQ/FAQformat.html#format1
482 https://github.com/igvteam/igv/wiki/TDF-Format

483 **Evaluating phasing completeness with parental genomes**

484 *Haplotype-specific k-mers (hap-mers)*

485 Parental haplotype markers can be obtained directly from the parental or ancestral
486 genomes[10]. In brief, distinct k-mers found in only one parent are collected, and the erroneous low-
487 frequency k-mers are filtered out. This filtering strategy relies on the k-mer count histogram, where
488 the cutoff for identifying reliable k-mers is computed as described above. When the child's short-
489 read data is also available, the inherited haplotype-specific markers can be obtained by

490 intersecting the child's k-mers with the parental marker sets. This time, we keep the k-mer counts
491 from the child's reads for further ploidy estimation (**Fig 3**).

492 *Hap-mer blob plot*

493 To get a global visual representation of the phasing completeness on assembled sequences,
494 we can count the total number of hap-mers found on each contig or scaffold (**Fig 4c-d**). Here,
495 each axis becomes the number of hap-mers found in a sequence entry (contig or scaffold). Each
496 circle (blob) represents a sequence, the size being relative to the length of the sequence.
497 Sequence bubbles near the diagonal represent mixtures of markers from both haplotypes, while
498 bubbles closer to a haplotype axis are predominately comprised of markers from a single
499 haplotype.

500 *Phase block statistics and switch error*

501 Whenever a hap-mer switch occurs, Merqury can flag that position in the assembly and output
502 a haplotype block report. This feature is useful for identifying phase blocks from a partially or
503 completely phased assembly such as FALCON-Unzip[9], FALCON-Phase[12], or Supernova2[4].
504 Merqury defines a phase block as a continuous sequence with at least two hap-mers originating
505 from the same haplotype (**Fig. 4a**). Usually, short-range switches are caused by consensus,
506 rather than phasing, errors. By default, Merqury allows at most 100 hap-mer switches per 20 kbp
507 as short range switches within a phase block. Each unexpected hap-mer found will be counted
508 as a switch error, but will not necessarily terminate the phase block. Ideally, when no switches
509 are found, the phase block N50 will be the same as the scaffold N50 and the sum of the phase
510 blocks will be identical to the assembled sequences. In reality, a scaffold often does not end with
511 a haplotype specific sequence, so the total phase block length is shorter. Trio-binned or
512 haplotype-resolved assemblies are a special case, where the entire haplotype assembly is
513 essentially a single block. Still, in this case, hap-mers from the other haplotype can be counted

514 as switch errors. Merqury also provides an option to restrict phase blocks to contigs and break
515 the blocks at any gap. At the end, Merqury reports the number of switches and total hap-mers on
516 each block along with the switch error rate in order to identify blocks with more frequent switch
517 errors.

518 *Hap-mer copy number spectrum*

519 The total k-mer spectrum of the assembled individual is also useful for tracking the fates of
520 haplotype-specific k-mers in a diploid assembly. Similar to the overall copy number analysis
521 performed with spectra-cn plots, we can count k-mers in each haplotype assembly and estimate
522 the completeness of haplotype-specific assembled bases compared to the hap-mer sets. For
523 example, by plotting separate histograms for hap-mers of different copy numbers in the assembly
524 (**Fig 6**), we can identify whether the assembly is artificially collapsing or duplicating sequence in
525 each haplotype. If hap-mers appear over (or under) represented by the assembly relative to the
526 read set, it is an indicator of artifactual duplication (or absence) of haplotype-specific sequence.
527 When evaluating a pseudo-haplotype assembly, which is designed to collapse or pick one
528 haplotype as much as possible, we can count the number of hap-mers present in the child's read
529 set but not present in the pseudo-haplotype assembly and use this to quantify the amount of
530 under-represented haplotype sequence. These missing hap-mers could then be used to identify
531 a set of alternative haplotype reads that were incorrectly excluded from the assembly.

532 **Assemblies**

533 All TrioCanu assemblies were downloaded from Koren *et al.*[10], available at
534 https://obj.umiacs.umd.edu/marbl_publications/triobinning/index.html. The *A. thaliana* F1
535 FALCON-Unzip assembly was obtained from Chin *et al*[9]. We generated a Canu assembly to show
536 a typical example of an interim mixed-haplotype assembly that has not been polished or purged

537 of haplotypic duplications. The same Pacific Bioscience reads was used for all three assemblies,
538 obtained from Chin *et al*[9].

539 The Canu assembly was generated with Canu 1.9 release version using the following
540 command:

```
541 canu -p canu -d athalF1_notrio genomeSize=130m 'corOutCoverage=100'
542 'batOptions=-dg 6 -db 6 -dr 1 -ca 500 -cp 50' 'batMemory=200' -pacbio-
543 raw *.fastq.gz
```

544 **Haplotype specific k-mers (hap-mers) for *A. thaliana F1* and NA12878**

545 Appropriate size of k was obtained as k=18 for the *A. thaliana* F1 with 130~260 Mbp genome
546 size and k=21 for NA12878 with 3.2~6.4 Gbp genome size using `$MERQURY/best_k.sh`.

547 As parental Illumina sequencing was not available for this F1, the parental genome
548 assemblies from Chin *et al*[9]. were used to obtain parental specific k-mers. Each assembly from
549 the inbred Col-0 and Cvi-0 lines were downloaded from:

550 https://downloads.pacbcloud.com/public/dataset/PhasedDiploidAsmPaperData/FUNZIP-
551 PhasedDiploidAssemblies.tgz.

552 Meryl databases for the parental strains were built directly with `meryl count k=18 output`
553 `$hap.meryl $hap.fasta` for each haplotype assembly.

554 The parental Illumina whole-genome sequencing sets for NA12878 were downloaded from
555 the high coverage dataset of the 1000 Genomes Project (NA12891 and NA12892) and combined
556 with Illumina Platinum Genomes Project data from PRJEB3381. Illumina whole-genome
557 sequencing of NA12878 was downloaded from PRJEB3381.

558 All Meryl databases from sequencing read sets were built with
559 `$MERQURY/_submit_build.sh`. Once the k-mer databases were built, inherited hap-mers
560 were obtained with `$MERQURY/trio/hapmers.sh`.

561 **Merqury on all assemblies**

562 Merqury was run for the *A. thaliana* F1 and NA12878 with the following command line for
563 diploid assemblies, where `$hap1` and `$hap2` are maternal (mat) and paternal (pat) for the
564 TrioCanu assemblies, and primary (pri) and alternate (alt) for FALCON-Unzip assemblies.

```
565 $MERQURY/_submit_merqury.sh $sample.k21.meryl mat.inherited.meryl
566 pat.inherited.meryl $hap1.fasta $hap2.fasta $asm_name
```

567 The *A. thaliana* F1 Canu assembly was run with:

```
568 $MERQURY/_submit_merqury.sh $sample.k21.meryl mat.inherited.meryl
569 pat.inherited.meryl canu.contigs.fasta Canu
```

570 **BUSCO**

571 BUSCO v3 was run on using `embryophyta_odb9` for the *A. thaliana* Canu assembly and
572 the combined Col and Cvi TrioCanu assembly with the following commands:

```
573 python run_BUSCO.py -i asm.fasta -o SAMPLE -l embryophyta_odb9 -m
574 genome -c 16 -sp arabidopsis
```

575 For NA12878, BUSCO was run in the same way, using `mammalia_odb9` for the combined
576 TrioCanu maternal and paternal assembly. BUSCO scores for each haplotype of *A. thaliana F1*
577 and NA12878 were obtained from Koren et al.[10] Supplementary Table 2.

578

QV estimates

CPU Time, memory consumption, and disk usage was measured for generating QVs on each haploid assemblies and the combined diploid assembly. A Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz node was used allowing up to 24 CPUs. Detailed node information is available at https://hpc.nih.gov.

Meryl-based QV

Meryl-based QV estimation to benchmark computing resources was evaluated for the counting (count), merging (union-sum), and QV steps with the following command:

```
$MERQURY/eval/qv.sh NA12878.k21.meryl mat.fasta pat.fasta meryl_qv
```

This generates Meryl databases for mat.fasta and pat.fasta, does a union for the databases, and generates QV scores for all three combinations (maternal, paternal, and both).

Mash-based QV

Mash-based QV estimation was performed using sketch size of `1000000` with the same k-mer size of 21:

```
mash sketch -s 1000000 -k $k $asm
mash screen -p $cpus $asm.msh `cat $input_fofn | tr '\n' ' '` >
$name.msh.idy
cat $name.msh.idy  | awk -v name=$name '{print name"\t"$2"\t"-
10*log(1-$1)/log(10)"\t"(1-$1)}' | tr '/' '\t' > $name.msh.qv
```

Mapping-based QV estimates

The Illumina WGS reads used to build the Meryl database were aligned to both mat.fasta and pat.fasta using BWA[45]. Base pair errors were called using FreeBayes[46] v1.3.1 `--skip-coverage 600`, skipping variant calling in regions with >600x read depth to help prevent unnecessary computing on high coverage regions that violate the diploid assumption. The exact commands used were:

```
# Add mat and pat at sequence names to prevent naming collisions in contigs
sed ‘s/>/>mat_h_/g’ mom.fasta > mat.fasta
sed ‘s/>/>pat_h_/g’ dad.fasta > pat.fasta

# bwa indexing
bwa index both.fasta

# bwa alignment
bwa mem -t 24 both.fasta F_1.fastq.gz F_2.fastq.gz > na12878.sam

# Sorting and converting to bam
samtools sort -@24 -O bam -o na12878.bam -T na12878.tmp na12878.sam

# Freebayes variant calling
freebayes --bam na12878.bam --skip-coverage 600 -f both.fasta |
bcftools view --no-version -Ou > na12878.tmp.bcf
bcftools index na12878.tmp.bcf

# Normalize indels
bcftools view -Ou -e'type="ref"' na12878.tmp.bcf | bcftools norm -Ob -f both.fasta -o na12878.bcf --threads 24
bcftools index na12878.bcf

# Filter out low quality variant calls
bcftools view -i 'QUAL>1 && (GT="AA" || GT="Aa")' -Oz --threads=24 na12878.bcf > na12878.changes.vcf.gz
```

```
bcftools index na12878.changes.vcf.gz

# Get number of bases called as errors
bcftools view -H -i 'QUAL>1 && (GT="AA" || GT="Aa")' -Ov
na12878.changes.vcf.gz | awk -F "\t" '{print $4"\t"$5}' | awk
'{lenA=length($1); lenB=length($2); if (lenA < lenB ) {sum+=lenB-
lenA} else if ( lenA > lenB ) { sum+=lenA-lenB } else {sum+=lenA}}
END {print sum}' > na12878.numvar

# Get number of bases with alignment
samtools view -F 0x100 -u na12878.bam | bedtools genomecov -ibam -
-split > aligned.genomecov
awk -v l=3 -v h=600 '{if ($1=="genome" && $2>l && $2<h) {numbp +=
$3}} END {print numbp}' aligned.genomecov > na12878.numbp

# QV calculation
NUM_BP=`cat na12878.numbp`
NUM_VAR=`cat na12878.numvar`
QV=`echo "$NUM_VAR $NUM_BP" | awk '{print (-
10*log($1/$2)/log(10))}'`
echo $QV
```

This pipeline is used in the Vertebrate Genomes Project, and the code used is available from https://github.com/VGP/vgp-assembly/tree/master/pipeline/ under the “bwa”, “freebayes-polish”, and “qv” directories.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: Merqury is openly available from https://github.com/marbl/merqury. Meryl is available from https://github.com/marbl/meryl. Meryl v1.0 release was used in this manuscript. Downloadable links to the Meryl k-mer databases of the assembled individuals and hap-mers are available for *A. thaliana* and NA12878 on the Merqury GitHub page.

Competing interests: SK has received travel compensation to present at Oxford Nanopore *meetings.* All other authors have no competing interests to declare.

Funding: This research was supported in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

Authors' contributions: A.R. conceived the project and implemented Merqury. B.P.W. designed and implemented Meryl. A.R., B.P.W., S.K., and A.M.P. designed the methods. A.R. and S.K. performed analyses. A.R. and A.M.P. coordinated the project. All of the authors wrote the manuscript and approved the final manuscript.

Acknowledgements: The authors would like to thank the members of the Vertebrate Genomes Project assembly working group for helpful discussions, including Marcela Uliano da Silva and Giulio Formenti. We also would like to thank Bernardo Clavijo of the Earlham Institute for his very insightful comments on a preprint of this manuscript. This research utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov).

References

1. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
2. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
3. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
4. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* (2017) doi:10.1101/gr.214874.116.
5. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
6. Falconer, E. *et al.* DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–1112 (2012).
7. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
8. Seo, J.-S. *et al.* De novo assembly and phasing of a Korean human genome. *Nature* **538**, 243–247 (2016).
9. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
10. Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182 (2018).
11. Ghareghani, M. *et al.* Strand-seq enables reliable separation of long reads by chromosome via expectation maximization. *Bioinformatics* **34**, i115–i123 (2018).
12. Kronenberg, Z. N. *et al. Extended haplotype phasing of* de novo *genome assemblies with FALCON-Phase*. http://biorxiv.org/lookup/doi/10.1101/327064 (2018) doi:10.1101/327064.
13. Garg, S. *et al.* Efficient chromosome-scale haplotype-resolved assembly of human

genomes. *bioRxiv* 810341 (2019) doi:10.1101/810341.

14. Porubsky, D. *et al. A fully phased accurate assembly of an individual human genome*. http://biorxiv.org/lookup/doi/10.1101/855049 (2019) doi:10.1101/855049.
15. Bickhart, D. M. *et al.* Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **49**, 643–650 (2017).
16. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinforma. Oxf. Engl.* **31**, 3210–3212 (2015).
17. 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).
18. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinforma. Oxf. Engl.* **34**, i142–i150 (2018).
19. Choi, Y., Chan, A. P., Kirkness, E., Telenti, A. & Schork, N. J. Comparison of phasing strategies for whole human genomes. *PLOS Genet.* **14**, e1007308 (2018).
20. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1–16 (2019).
21. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2017).
22. Patterson, M. *et al.* WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **22**, 498–509 (2015).
23. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
24. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 1–26 (2016).

25. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).
26. Eberle, M. A. *et al.* A 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. Miga, K. H. *et al. Telomere-to-telomere assembly of a complete human X chromosome*. http://biorxiv.org/lookup/doi/10.1101/735928 (2019) doi:10.1101/735928.
28. 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).
29. Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nat. Rev. Genet.* 1–12 (2020) doi:10.1038/s41576-020-0210-7.
30. Human Pangenome Reference Consortium. *Human Pangenome Reference Consortium* https://humanpangenome.org/.
31. Vertebrate Genomes Project. *Vertebrate Genomes Project* https://vertebrategenomesproject.org.
32. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* gr.215087.116 (2017) doi:10.1101/gr.215087.116.
33. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
34. Guan, D. *et al. Identifying and removing haplotypic duplication in primary genome assemblies*. *Bioinformatics* doi:10.1093/bioinformatics/btaa025 (2020).
35. Vurture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
36. Robinson, J. T. *et al.* Integrative Genomics Viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
37. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV):

high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

38. Ondov, B. D. *et al.* Mash Screen: high-throughput sequence containment estimation for genome discovery. *Genome Biol.* **20**, 232 (2019).
39. Quail, M. A. et al. A large genome center's improvements to the Illumina sequencing system. Nat. Methods **5**, 1005–1010 (2008).
40. Kozarewa, I. et al. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat. Methods **6**, 291–295 (2009).
41. Oyola, S. O. et al. Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. BMC Genomics **13**, 1 (2012).
42. Miller, J. R. et al. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics **24**, 2818–2824 (2008).
43. Fofanov, Y. et al. How independent are the appearances of n-mers in different genomes? Bioinformatics **20**, 2421–2428 (2004).
44. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
45. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013).
46. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *ArXiv12073907 Q-Bio* (2012).