PROF. HOLGER SCHIELZETH (Orcid ID : 0000-0002-9124-2261)

DR. PHILIP EWELS (Orcid ID : 0000-0003-4101-2502)

DR. JESÚS T. GARCÍA (Orcid ID : 0000-0003-4126-9658)

DR. ALEXANDER SUH (Orcid ID : 0000-0002-8979-9992)

DR. RETO BURRI (Orcid ID : 0000-0002-1813-0079)

Article type : Resource Article

Molecular and Statistical Advances

Linked-read sequencing enables haplotype-resolved resequencing at population scale

Dave Lutgen[1*] | Raphael Ritter[1*] | Remi-André Olsen[2] | Holger Schielzeth[1] | Joel Gruselius [3] | Phil Ewels[2] | Jesús T. García[4] | Hadoram Shirihai[5] | Manuel Schweizer[5,6] | Alexander Suh[7,8] |Reto Burri[1]

[1] Department of Population Ecology, Institute of Ecology and Evolution, Friedrich Schiller University Jena, Dornburger Strasse 159, D-07743 Jena, Germany

[2] Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, Box 1031, SE-17121 Solna, Sweden

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1755-0998.13192

[3] Science for Life Laboratory, Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden

[4] Instituto de Investigación en Recursos Cinegéticos (IREC), CSIC-UCLM-JCCM, Ronda de Toledo 12, 13005 Ciudad Real, Spain

[5] Natural History Museum Bern, Bernastrasse 15, CH-3005 Bern, Switzerland

[6] Institute of Ecology and Evolution, University of Bern, CH-3012 Bern, Switzerland

[7] Department of Organismal Biology – Systematic Biology, Evolutionary Biology Centre (EBC), Uppsala University, Norbyvägen 18D, SE-75236 Uppsala, Sweden

[8] Present address: School of Biological Sciences, University of East Anglia, Norwich Research Park, NR4 7TU, Norwich, United Kingdom

* These authors contributed equally to the work

Correspondence
Reto Burri, burri@wildlight.ch

Short title: Haplotype-resolved population resequencing

Keywords: admixture, demography, introgression, phasing, population genomics, selective sweeps

Abstract

The feasibility to sequence entire genomes of virtually any organism provides unprecedented insights into the evolutionary history of populations and species. Nevertheless, many population genomic inferences – including the quantification and dating of admixture, introgression and demographic events, and inference of selective sweeps – are still limited by the lack of high-quality haplotype information. The

newest generation of sequencing technology now promises significant progress. To establish the feasibility of haplotype-resolved genome resequencing at population scale, we investigated properties of linked-read sequencing data of songbirds of the genus *Oenanthe* across a range of sequencing depths. Our results based on the comparison of downsampled (25x, 20x, 15x, 10x, 7x, and 5x) with high-coverage data (46-68x) of seven bird genomes mapped to a reference suggest that phasing contiguities and accuracies adequate for most population genomic analyses can be reached already with moderate sequencing effort. At 15x coverage, phased haplotypes span about 90% of the genome assembly, with 50 and 90 percent of phased sequences located in phase blocks longer than 1.25-4.6 Mb (N50) and 0.27-0.72 Mb (N90). Phasing accuracy reaches beyond 99% starting from 15x coverage. Higher coverages yielded higher contiguities (up to about 7 Mb/1Mb (N50/N90) at 25x coverage), but only marginally improved phasing accuracy. Phase block contiguity improved with input DNA molecule length; thus, higher-quality DNA may help keeping sequencing costs at bay. In conclusion, even for organisms with gigabase-sized genomes like birds, linked-read sequencing at moderate depth opens an affordable avenue towards haplotype-resolved genome resequencing at population scale.

Introduction

The possibility to sequence entire genomes at population-scale has nothing short of revolutionized molecular ecology. Genome-wide data are now routinely used to infer populations' histories of demography, admixture, and introgression (Beichman *et al.* 2018; Gompert & Buerkle 2013; Taylor & Larson 2019); investigations of genetic diversity along genomes have provided unprecedented insights into the distribution of most differentiated regions across genomes (e.g. Burri *et al.* 2015; Jones *et al.* 2012; Martin *et al.* 2013; Soria-Carrasco *et al.* 2014) and the processes driving their evolution (reviewed e.g. in Burri 2017; Martin & Jiggins 2017; Ravinet *et al.* 2017); and numerous phenotypes have been mapped to genomic regions underpinning their expression (e.g. Kupper *et al.* 2016; Lamichhaney *et al.* 2015; Santure & Garant 2018; Schielzeth *et al.* 2018). Enabling these insights, studies of genome-wide variation continue contributing to our understanding of the evolutionary histories of phenotypes, populations, and species at an ever increasing pace.

Nonetheless, deeper insights into the molecular ecology of natural populations – in particular the extent and timing of admixture and introgression, and the distribution of genomic regions that underwent selective sweeps – can be achieved by the integration of yet usually difficult to obtain haplotype information, that is, information on the gametic phase of genetic variants (e.g. Palamara *et al.* 2012). Haplotype structure reflects some of the most significant footprints left by admixture (Buerkle & Rieseberg 2008; Fisher 1949, 1954; Pool & Nielsen 2009), selective sweeps (e.g. Sabeti *et al.* 2002; Voight *et al.* 2006), or a combination thereof (e.g. Shchur *et al.* 2019). Haplotype tracts that enter populations (according to context referred to as 'migrant tracts' or 'ancestry tracts') are progressively broken down to smaller size as recombination events accumulate over evolutionary time (e.g. Janzen *et al.* 2018; Pool & Nielsen 2009). According to such a recombination clock (Moorjani *et al.* 2016), long haplotype tracts are of more recent origin than shorter ones. Thus, migrant tracts that entered populations/species through admixture/ introgression recently are expected to be longer than migrant tracts of older origin (Pool & Nielsen 2009). The size distribution of migrant tracts thus enables powerful insights into histories of admixture and introgression (e.g. Palamara & Pe'er 2013). Likewise, long haplotype tracts that reached high frequency in a population must have done so in a timeframe that did not allow recombination to break them down – thus fast and likely by positive selection (e.g. Sabeti *et al.* 2002; Voight *et al.* 2006). Whereas current inference methods that predominantly rely on allelic states and/or allele frequencies may be limited in their power to distinguish among complex demographic hypotheses and detecting

selective sweeps, these tasks may be greatly facilitated by haplotype-based approaches (Gompert & Buerkle 2013).

However, the application of haplotype-based approaches to natural populations has yet been stalled by the difficulty to obtain phased genomic data, that is, data for which the gametic phase of genetic variants is known. Although gametic phase can be inferred statistically (e.g. Delaneau *et al.* 2012; Stephens *et al.* 2001), this typically requires at least 100 individuals to obtain solid phase information (Browning & Browning 2011), surmounting what sequencing experiments in natural populations can usually achieve. With a lower number of individuals, statistical phasing has low accuracy and reduced power (Browning & Browning 2011; Choi *et al.* 2018), and may limit for instance the inference of selective sweeps (Nadachowska-Brzyska *et al.* 2019). Therefore, haplotype-based approaches to infer admixture and selection have largely been limited to humans, where thousands of available genome sequences enable statistical phasing with reasonable confidence (e.g. Loh *et al.* 2016; O'Connell *et al.* 2016). A notable exception is a recent study on natural populations of sea bass that used parent-offspring trios to accurately infer the gametic phases of genome-wide data (Duranton *et al.* 2018). However, this approach is expensive, laborious, and not applicable to all species. Alternative approaches to obtain phased genomic data are thus called for.

The latest generation of genome sequencing technologies now offers promising avenues towards cost-efficient haplotype-resolved genome sequencing – sequencing that directly determines gametic phase from template DNA molecules (Snyder *et al.* 2015). Long-read technologies, such as offered by Pacific Biosciences (PacBio) and Oxford Nanopore, determine genome sequences from single molecules. These approaches yield sequences with lengths up to tens of kilobases (kb), and accordingly provide phase information across at least the same scale. Still, both methods require high quantities of input DNA and high quality in terms of molecule length, and remain prohibitively expensive for population-scale sequencing experiments. On the other hand, affordable standard short-read sequencing by itself provides phasing over at most a couple of hundred base pairs. However, dedicated 'linked-read' library preparation protocols (**Figure 1**) now enable haplotype-resolved sequencing based on short-read sequencing (Chen *et al.* 2019; Redin *et al.* 2019; Snyder *et al.* 2015; Wang *et al.* 2019; Zheng *et al.* 2016) that outperforms statistical phasing of human data (Choi *et al.* 2018). One approach to the preparation of such linked-read performs library preparation in emulsion droplets that each contain a very limited number of DNA molecules. The addition of droplet-specific barcodes during library preparation distinguishes short reads from different DNA molecules and thereby enables efficient resolution of haplotypes of lengths up to over hundred kilobases (Weisenfeld *et al.* 2017; Zheng *et al.* 2016). Alternatively, transposases are used to

introduce barcodes that distinguish sequence fragments originating from different DNA molecules (Chen *et al.* 2019; Wang *et al.* 2019). Thus, while for the time being long-read sequencing of multiple individuals remains prohibitively expensive (but see Weissensteiner *et al.* 2019), on top of enabling cost-efficient and contiguous *de novo* genome assemblies of gigabase-sized genomes (e.g. Boman *et al.* 2019; Kinsella *et al.* 2019; Schweizer *et al.* 2019a; Toomey *et al.* 2018), linked-read sequencing opens the scope for haplotype-resolved sequencing at population scale.

Here, we set out to determine the linked-read sequencing effort required to obtain accurate and long-range phase information. Such information can help maximizing the cost-efficiency of linked-read sequencing for haplotype-resolved sequencing at population scale. To this end, we performed a downsampling experiment on seven about 1 gigabase-sized bird genomes originally sequenced at 48-68x coverages, and determined phasing contiguities and accuracies at coverages of 25x down to 5x. Our results suggest that adequate phase block contiguities and phasing accuracies can be reached with read-depths as low as 15x. This implies that haplotype-resolved sequencing is feasible at population scale, and opens new perspectives for the implementation of high-quality phase information in demographic reconstructions, including histories of hybridization and introgression (Harris & Nielsen 2013; Lawson *et al.* 2012; Palamara & Pe'er 2013), genome scans for selective sweeps and reproductive incompatibilities (Ferrer-Admetlla *et al.* 2014; Sabeti *et al.* 2002; Sedghifar *et al.* 2016; Tang *et al.* 2007; Voight *et al.* 2006), and conservation genomics (Duranton *et al.* 2019; Leitwein *et al.*).

Materials & Methods

DNA extraction, library preparation, and genome sequencing

We sequenced the genomes of seven individuals of the four species in the *Oenanthe hispanica-pleschanka-melanoleuca-cypriaca* complex (Schweizer *et al.* 2019a) within the open-habitat chats group of songbirds (Aliabadian *et al.* 2012; Schweizer *et al.* 2019b) (**Table 1**). To this end, DNA was extracted from blood or tissue conserved in ethanol from each two western black-eared wheatears (*O. hispanica*), two pied wheatears (*O. pleschanka*), two eastern black-eared wheatears (*O. melanoleuca*), and a Cyprus wheatear (*O. cypriaca*) (**Table 1**) using the MagAttract HMW DNA kit (Qiagen, Hilden, Germany). Digestion was performed in 360 ul of buffer ATL, 440 ul of buffer AL, and 20 ul of proteinase K starting at 56°C for one hour and 37°C overnight. Additional 20 ul of proteinase K were added in the morning, and digestion completed at 56°C. Total time of digestion was about 16 hours. DNA extraction then followed the manufacturer's recommendations.

Linked-read sequencing libraries were prepared using Chromium Genome library kits (10X Genomics) and each library was sequenced on half a lane of an Illumina HiSeq X flowcell. Distributions of the numbers of reads per 10X Genomic barcode are provided in **Figure S1**.

Reference genome draft assembly and curation

We assembled the genomes of all seven individuals using the Supernova 2.1 assembler (a significantly less contiguous Supernova 1 assembly of *O. melanoleuca* YPM 101348 was published in Schweizer *et al.* 2019a and is available from Dryad doi:10.5061/dryad.6d006j3). The genome of the Cyprus wheatear (*O. cypriaca*; **Table 1**) was chosen as reference assembly for all downstream analyses as it had highest scaffold N50. The merged, pseudohaploid assembly was 1.08 Gb long with a scaffold N50 of 25.15 Mb in 23,105 scaffolds. To assess assembly completeness, we evaluated the presence, completeness, and copy number of avian benchmarking universal single-copy orthologs (BUSCOs, aves_odb9, creation date: 2016/02/13) as assessed by BUSCO version 3 (Simão *et al.* 2015). Of 4,915 BUSCOS, 4,462 (90.8%) were complete and single copy, 62 (1.3%) were complete and duplicated, 214 (4.4%) were fragmented, and 177 (3.6%) were missing.

To remove duplicate scaffolds of at least 99% identity, we ran the dedupe procedure in BBTools (https://sourceforge.net/projects/bbmap/) allowing up to 7,000 edits. This reduced the assembly to 11,030 scaffolds. We then aimed to ensure that all duplicate scaffolds were removed and retain only scaffolds whose integrity can be confirmed by the presence of syntenic regions in another songbird genome. To this end, we performed a lastz alignment against the collared flycatcher assembly version 1.5 (Kawakami *et al.* 2014), which is the highest-quality assembly available from the Muscicapidae family. For this we used lastz 1.04 (Harris 2007) with settings M=254, K=4500, L=3000, Y=15000, C=2, T=2, and --matchcount=10000. This resulted in 295 scaffolds with unique hits in the flycatcher assembly. The final assembly spanned 970 Mb, with a scaffold N50 of 32.13 Mb. All following analyses used this Cyprus wheatear draft genome assembly as reference.

Evaluation of phasing accuracy, genotyping accuracy, and phase set contiguity

To obtain proxies for phasing and genotyping accuracy, we assumed the phasing and genotyping of the full data for each individual to be correct (46-68x coverage, **Table 1**). Mismatches in phase and genotype between downsampled and full data were considered as inaccurate phasing and genotyping, respectively. These 'truth sets' may be less accurate than those based on parent-offspring or pedigree information, yet, unless phasing is systematically biased towards the same erroneous phase for both the downsampled and the full data, we do not expect phasing accuracy to be vastly overestimated. To evaluate these accuracies at different coverages, we downsampled the full data to 25x, 20x, 15x, 10x, 7x, and 5x average coverage directly in the LongRanger pipeline, which was also used for SNP calling and genotyping, and phasing (Longranger in GATK mode with GATK version 3.8.0, McKenna *et al.* 2010). In brief, to genotype and phase

SNPs, LongRanger first uses the Lariat aligner (Bishara *et al.* 2015) that maps reads to the reference genome using bwa (Li & Durbin 2009) and then applies linked-read information to resolve ambiguous mappings. It then marks duplicates using picard (https://github.com/broadinstitute/picard) and calls and genotypes SNPs using GATK HaplotypeCaller (McKenna *et al.* 2010). The linked-read information is applied to resolve the gametic phase. One-fold genome size for the downsampling procedure was assumed 1.08 Gb, which corresponds to the Supernova version 2.1 merged/pseudohaploid assembly length.

To determine whether phasing corresponded between the downsampled and full data sets, we parsed the genotypes including phasing information for SNPs that fulfilled the following criteria: 1) biallelic, heterozygous SNPs, which 2) passed LongRanger filter criteria, and 3) for which the genotype was identical between the downsampled and full data set (see also genotyping accuracy below). The number and length of phase sets (that is the set of SNPs that can be attributed to the same haplotype) differ between the full and the downsampled data set. Moreover, the order in which maternal or paternal haplotypes are output can differ between the full and downsampled data and needs to be matched before further analysis. To this end, we determined all combinations of phase sets between the full and downsampled data before estimating phasing mismatches. For each phase set combination we then determined whether phasing was resolved in the same order in the two data sets or whether it had to be inversed for phasing to be compared. For phase set combinations for which more phasings were wrong than correct if retained in the original order, we inversed phasing order before determining phasing accuracy.

Genotyping accuracies were inferred as the percentage of SNPs with matching genotype between the downsampled and full data sets (before applying the three criteria above).

Phase block contiguity was estimated in terms of phase set N50 and N90. These values were estimated in reference to the combined lengths of all phase sets. That is, phase sets equal and longer than N50/N90 cover 50/90 percent of the number of base pairs contained in all phase sets combined. Furthermore, we estimated the proportion of the genome that was spanned by phase sets. We estimated this proportion as the sum of phase set lengths divided by the length of the genome assembly. It therefore reflects the fraction of the genome for which any kind of phasing information is available without addressing its completeness.

Scripts used to evaluate LongRanger outputs are provided as Supplemental Information. All analyses were performed on data sets that alternatively contained unfiltered SNPs (at least one read) or SNPs for which at least 5, 8, 10, or 12 reads were available.

Statistical analysis

To determine the effect of sequencing coverage, SNP filtering and input DNA molecule length on phasing and genotyping accuracy and on phase block contiguity, we ran linear models in R 3.6.3 (R Core Team 2020). N50 and N90 were log-transformed, accuracies transformed as -log(100-x), and explanatory variables centred. We tested models including all variables but retained only significant variables in the final models, starting removal of non-significant terms by their P-values and retaining main effects that are involved in interactions. Because comparable estimates of physical DNA fragment size distributions were available only for five samples, we determined input DNA molecule length for each sample in Supernova version 2.1. For the five samples for which estimates of peak DNA fragment sizes were available from Fragment Analyzer runs (DNF-464-33 - HS Large Fragment 50 kb), these strongly correlated with the input DNA molecule lengths estimated by Supernova 2.1 (Pearson's correlation, 0.91; p=0.033). Supernova 2.1 estimates were on average 1.39 ± 0.28 (mean ± standard deviation) times higher than peak fragment sizes estimated on Fragment Analyzer. Analyses on the reduced data set for which peak fragment lengths estimated with Fragment Analyzer were available yielded congruent results and are not reported.

Results

Phase block contiguity

Phase block contiguity increased by one to two orders of magnitude with increasing sequencing coverage (by at most five times), and was significantly higher for longer input DNA molecule length (**Figure 2A-B,D-E**, **Table 2**).

Contiguity in terms of phasing N50 (and N90) increased from about 10-530 kb (1-140 kb) at 5x coverage to 1,390-7,120 kb (190-1,020 kb) at 25x coverage (**Figure 2A-B**). As shown by N50/N90 for full data, for input DNA >30 kb contiguity continues increasing at higher coverages (**Figure 2D-E**). At a given coverage, contiguity increased largely linearly with increasing input molecule length (**Figure S2**). Filtering for a minimum read depth per SNP had no detectable effect on contiguity (**Figure S3**, p=0.92). The effect of molecule length increased with coverage, as indicated by their significant interaction (**Table 2**). Specifically, the gain in N50 increased from ca 8.5 kb per kb molecule length at 5x coverage to almost 90 kb per kb molecule length at 25x (**Figure 3**). Molecule length seemed a better predictor of phase block contiguity at low and high than at intermediate coverages (see larger residuals for the latter in **Figure S2**).

Coverage of the genome by phase sets

Almost 90% of the genome assembly was covered by phase sets (that is sets of SNPs that are combined into haplotypes) provided adequate input DNA molecule length and a read-depth filtering of SNPs not exceeding average coverage (mean ± standard deviation: 87.1 ± 1.8 %, **Figure 2C,F**, **Table 3**). The proportion of the genome covered by phase sets increased with coverage and input molecule length and decreased with filtering stringencies that exceed sequencing coverage (**Figure 2C,F**, **Table 3**). With input molecule lengths >30 kb, this proportion plateaued slightly below 90% for coverages starting at 15x, independent of filtering (**Figure S3**). Read-depth filters exceeding sequencing coverage drastically reduced the proportion of the genome covered by phase sets. This is expected, as such filtering simply removes most SNPs from the data set.

Phasing and genotyping accuracy

Both phasing and genotyping accuracies strongly increased with increasing coverage and with adequate read-depth filtering (**Figure 4**, **Table 4**).Phasing accuracy increased from down to almost 60% at 5x coverage to above 99% starting at 15x coverage for input DNA with molecule lengths estimated >30 kb (**Figure 4**, **Table 4**). Samples with input DNA <20 kb still reached phasing accuracies >98% (starting from 15x coverage). Filtering for a minimum read-depth improved phasing accuracy significantly at low coverages (**Figure 4**, **Table 4**). At all but low coverages, phasing accuracy increased with increasing molecule length (**Figure S4**, **Table 4**; at low coverage phasing accuracy was lower for long molecule lengths). Genotyping accuracy likewise strongly increased with increasing coverage (**Figure 4**, **Table 4**) and was further improved by adequate read-depth filters. Accuracies above 99% were reached starting at a coverage of 15x with a read-depth filter of at least 8. However, overfiltering (for higher than genome-wide average coverage) strongly decreased genotyping accuracy (**Figure 4**), as indicated by a significant interaction of filtering with coverage (**Table 4**). Finally, at coverages of 15x or higher, genotyping accuracy increased with increasing input molecule length, while the opposite effect was observed at lower coverages (**Figure S5**).

Discussion

Our analyses suggest that phase block contiguity and accuracy, as well as genotyping accuracy of haplotype-resolved sequencing improve with increasing sequencing coverage and mean input DNA quality in terms of molecule length. Additional improvements of genotyping accuracy can be reached with

adequate read-depth filtering. However, with adequate input DNA (> 30 kb), phasing contiguities at Mb-scale with high phasing accuracies in the same range as genotyping accuracy can be achieved already with a sequencing coverage of 15x. Thereby, the sequencing effort required to obtain haplotype-resolved genome resequencing data of high quality is in the same range as currently used in many large-scale population resequencing studies (e.g. Burri *et al.* 2015; Martin *et al.* 2013; Stankowski *et al.* 2019; Vijay *et al.* 2016), and thus affordable both financially and in terms of the amount of input DNA required. Our study thus reveals an affordable avenue towards haplotype-resolved genome resequencing data at population scale, even for organisms with gigabase-sized genomes like birds. In the following, we discuss the implications of our results and the recommendations deriving therefrom for the design of studies that aim to make use of haplotype-resolved sequencing.

Towards population-scale haplotype-resolved sequencing

Although sequencing at high coverage yields highest phase block contiguity, moderate coverages around 15x open the avenue for high-quality phasing at populations scale.

According to our results, best phase block contiguities and accuracies for each sample are achieved at high sequencing coverage (**Figures 2, 4**). Phase block contiguity reached N50 of >7 Mb at 25x coverage, and above this coverage continued increasing up to almost 11 Mb, as shown when adding contiguities of full data with coverages >45x (**Figure S3**). In contrast, phasing accuracy improved only marginally with coverages beyond 15x (**Figure 4**). From 15x coverage on, it exceeded 99% and was thus in the same range as genotyping accuracy. In contrast to genotyping accuracy, which can be strongly improved by filtering for a minimum read-depth, gains in phasing accuracy through filtering were moderate and risked the cost of covering a reduced proportion of the genome with phasing information (**Figure 4**). Higher than 15x sequencing coverage thus predominantly affords improved phasing contiguities rather than phasing accuracy.

However, our prime interest was in establishing whether adequate phasing contiguities and accuracies can be achieved with affordable sequencing effort to enable haplotype-resolved sequencing at population-scale. Indeed, our results suggest that already with reasonable sequencing effort phasing contiguities adequate for most applications in molecular ecology can be reached. At a coverage of 15x of linked reads, which is similar to the coverage nowadays used in many population-scale short-read sequencing experiments (e.g. Burri *et al.* 2015; Martin *et al.* 2013; Stankowski *et al.* 2019; Vijay *et al.* 2016), phasing N50 reaches Mb-scale (**Figure 2**) and phasing accuracy exceeding 99% (**Figure 4**). For samples with input DNA >30 kb, phasing accuracies at 15x and higher coverage are in the same range as

observed in equivalent human data, where linked-read data consistently outperformed statistical phasing in all respects (Choi *et al.* 2018). Moreover, the phasing accuracies estimated here compare to values reported for human genomes sequenced with the same technology at 25x coverage (<0.4%, Porubsky *et al.* 2017). Not even with parent-offspring trios current state-of-the-art statistical phasing approaches reach as high phasing accuracy as we find at this moderate coverage (Al Bkhetan *et al.* 2019).

Thus already sequencing coverages of 15x should enable high-quality inferences of phase blocks relevant to many population genetic inferences in outbred natural populations. In cichlids, for instance, average ancestry tract lengths is at most 3 kb; the majority of ancestry tract is smaller than 20 kb, and only few are longer than 50 kb (Meier *et al.* 2017). In European sea bass, even in low-recombination regions migrant tracts are on average 50 kb long; longest migrant tracts measure up to 2 Mb (Duranton *et al.* 2018). The longer ancestry tract sizes in sea bass versus cichlids reflects the more recent admixture in this system that started only about 11,500 years ago as opposed to up to 100,000-200,000 years ago in cichlids. Still, even for the case of relatively recent admixture in sea bass, phasing contiguities achievable with 15x coverage should span the majority of haplotype tracts. However, in systems with more recent admixture with onsets a couple of hundred years ago, such as between North American canids, ancestry tracts can be considerably longer, averages reaching several Mb (up to >9 Mb) (vonHoldt *et al.* 2011). In species with such a recent history of admixture, phasing contiguities achieved with 15x coverage indeed may be shorter than relevant haplotype tract lengths. However, additional statistical phasing of phase blocks with read-aware software that treats phase blocks as reads, such as Shapeit4 (Delaneau *et al.* 2019), may extend the phasing beyond relevant contiguities even for such recent cases of admixture.

In conclusion, our results suggest that with the same sequencing efforts as used in standard resequencing studies linked-read sequencing is able to provide valuable phasing information, thus paving the way for affordable haplotype-resolved genome resequencing at population scale.

Our data demonstrate the efficiency of linked-read sequencing for population genetic analyses. However, with the imminent demise of the 10X Genomics Chromium linked-read technology, alternative avenues towards population-scale haplotype-resolved resequencing are called for. Fortunately, solutions are in sight both from the newest generation of long-read sequencing and from alternative linked-read sequencing approaches. PacBio HiFi reads start reaching Illumina-level accuracy at costs that promise to soon drop to an affordable level (Nurk *et al.* 2020). Comparable accuracies may be reached with a hybrid approach combining highly accurate Illumina short-reads with phase information from Oxford Nanopore reads (Ebler *et al.* 2019). On the linked-read front, in particular transposase-based approaches, such as TELL-seq (Universal Sequencing, Chen *et al.* 2019) and Long-Fragment-Read Whole Genome Sequencing

(BGI, Wang *et al.* 2019) offer promising avenues. In contrast to long-read sequencing, linked-read solutions like these and the emulsion-based approach by 10X Genomics (**Figure 1**) are not limited by sequencing read length but chiefly by DNA quality. The near future will show, which of the available technologies offers the most affordable path towards best haplotype information. Universal computational tools for haplotype-reconstruction that can handle all types and combinations of data, such as hapCUT2 (Edge *et al.* 2017), may prove invaluable to this endeavour.

Issues with tissues: Phase block contiguity requires high-quality input DNA

Our results suggest that requirements for high phasing contiguities start with the choice of tissues for DNA extraction, or latest in the wet lab with adequate high molecular weight (HMW) DNA extraction methods. The physical size over which linked-read sequencing can provide haplotype information is in first line limited by the size of input DNA fragments. In line with this expectation, we observed highest phasing contiguities for DNA of high quality, that is with long input DNA molecules >30 kb; DNA of poor quality (<20 kb) did not yield good phase block contiguity. Moreover, for DNA of good quality the same phase block contiguity can be reached at low to moderate coverage as with poor DNA at high coverage (broken lines in **Figure 2**). Still, at intermediate coverages, the correlation of phase block contiguity with input molecule length is poorer, likely due to the stochasticity with which fragments are sequenced at these coverages. At these coverages the chance that parts of long molecules are not sequenced increases, and long molecules may be broken up into shorter phase sets. Nevertheless, best phasing contiguities require high-quality DNA. Therefore, tissue preservation and DNA extraction methods that yield HMW DNA help keeping sequencing costs affordable and may be key to take haplotype-resolved genome sequencing to population scale.

The importance of good DNA quality to obtain high phasing contiguities and keep sequencing costs at a minimum implies that measures to obtain the best possible source of DNA must be taken starting from tissue collection on. Systematic studies on the choice and preservation of tissue samples are yet lacking. Although our study was not designed to infer best practices for tissue preservation and DNA extraction for linked-read sequencing, our results suggest that DNA quality depends on the tissue of origin. We extracted DNA exclusively with a magnetic bead-based HMW DNA protocol, and neither was tissue age related to differences in DNA quality nor was there a difference in DNA preservation (all samples were stored in ethanol). The observed differences in molecule lengths thus likely reflect issues with the type of tissues used for DNA extraction. Indeed, the most notable characteristic of the two data sets with input molecule lengths <20 kb was the tissue of origin: in both cases this was muscle as opposed to blood for

the other samples (**Table 1**). Therefore, to preserve the best possible source for HMW bird DNA, we recommend that bird tissue collections include samples of blood (red blood cells are nucleated in birds). For museum-based cryo-collections, which often exclusively sample solid tissues from body cavities but not blood, this implies that blood be collected prior to sacrificing birds. After tissue collection, measures to protect DNA from damage are required, starting with the preservation in the lab/field and ending with DNA extraction. The suitability of different preservation buffers remains to be investigated. However, we recommend that samples be cooled at the best possible, and that DNA extraction makes use of methods dedicated to the isolation of HMW DNA (Klingström *et al.* 2018). To this end, if conducted with great care (that is without vortexing and pipetting up and down to mix), classical phenol-chloroform DNA extractions may yield DNA quality comparable to bead-based approaches. To avoid very long DNA strands from breaking and obtain ultra-HMW DNA, methods that support DNA molecules on magnetic platelets (offered e.g. by Circulomics) or in agarose gel-plugs (offered e.g. by Bionano Genomics) may be required.

While genomic studies have been limited by sequencing technology to date, sequencing now starts being limited by the quality of input DNA. We thus recommend that research in the wild today starts collecting tissues with qualities for tomorrow.

Acknowledgements

We are grateful to the Yale Peabody museum for providing tissue of a male *O. melanoleuca* (101348) and to the Museum of Vertebrate Zoology of Harvard University for providing tissue of a male *O. pleschanka* (349924). We thank Ana Gomes for extracting HWM DNA, and acknowledge support from the National Genomics Infrastructure in Stockholm funded by the Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Swedish Research Council, and the SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure. Further computation was performed at the High-Performance Computing Cluster EVE, a joint effort of the Helmholtz Centre for Environmental Research (UFZ) and the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig. We thank the administration and support staff of EVE: Thomas Schnicke and Ben Langenberg (UFZ), and Christian Krause (iDiv). This research was supported by a Science for Life Laboratory Swedish Biodiversity Program grant (2015-R14) to AS and by a German Research Foundation (DFG) research grant (BU3456/3-1) to RB.

References

Al Bkhetan Z, Zobel J, Kowalczyk A, Verspoor K, Goudey B (2019) Exploring effective approaches for haplotype block phasing. *BMC Bioinformatics* **20**, 540.

Aliabadian M, Kaboli M, Förschler MI*, et al.* (2012) Convergent evolution of morphological and ecological traits in the open-habitat chat complex (Aves, Muscicapidae: Saxicolinae). *Molecular Phylogenetics and Evolution* **65**, 35-45.

Beichman AC, Huerta-Sanchez E, Lohmueller KE (2018) Using Genomic Data to Infer Historic Population Dynamics of Nonmodel Organisms. *Annual Review of Ecology, Evolution, and Systematics* **49**, 433-456.

Bishara A, Liu Y, Weng Z*, et al.* (2015) Read clouds uncover variation in complex regions of the human genome. *Genome Research* **25**, 1570-1580.

Boman J, Frankl-Vilches C, da Silva dos Santos M*, et al.* (2019) The Genome of Blue-Capped Cordon-Bleu Uncovers Hidden Diversity of LTR Retrotransposons in Zebra Finch. *Genes* **10**, 301.

Browning SR, Browning BL (2011) Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics* **12**, 703-714.

Buerkle CA, Rieseberg LH (2008) The rate of genome stabilization in homoploid hybrid species. *Evolution* **62**, 266-275.

Burri R (2017) Interpreting differentiation landscapes in the light of long-term linked selection. *Evolution Letters* **1**, 118-131.

Burri R, Nater A, Kawakami T*, et al.* (2015) Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Research* **25**, 1656-1665.

Chen Z, Pham L, Wu T-C*, et al.* (2019) Ultra-low input single tube linked-read library method enables short-read NGS systems to generate highly accurate and economical long-range sequencing information for *de novo* genome assembly and haplotype phasing. *bioRxiv*, 852947.

Choi Y, Chan AP, Kirkness E, Telenti A, Schork NJ (2018) Comparison of phasing strategies for whole human genomes. *PloS Genetics* **14**, e1007308.

Delaneau O, Marchini J, Zagury J-F (2012) A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179-181.

Delaneau O, Zagury J-F, Robinson MR, Marchini JL, Dermitzakis ET (2019) Accurate, scalable and integrative haplotype estimation. *Nature Communications* **10**, 5436.

Duranton M, Allal F, Fraïsse C*, et al.* (2018) The origin and remolding of genomic islands of differentiation in the European sea bass. *Nature Communications* **9**, 2518.

Duranton M, Bonhomme F, Gagnaire P-A (2019) The spatial scale of dispersal revealed by admixture tracts. *Evolutionary Applications* **12**, 1743-1756.

Ebler J, Haukness M, Pesout T, Marschall T, Paten B (2019) Haplotype-aware diplotyping from noisy long reads. *Genome Biology* **20**, 116.

Edge P, Bafna V, Bansal V (2017) HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Research* **27**, 801-812.

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R (2014) On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. *Molecular Biology and Evolution* **31**, 1275-1291.

Fisher RA (1949) *The theory of inbreeding* Oliver and Boyd.

Fisher RA (1954) A fuller theory of "juctions" in inbreeding. *Heredity* **8**, 187-197.

Gompert Z, Buerkle CA (2013) Analyses of genetic ancestry enable key insights for molecular ecology. *Molecular Ecology* **22**, 5278-5294.

Harris K, Nielsen R (2013) Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PloS Genetics* **9**, e1003521.

Harris RS (2007) *Improved pairwise alignment of genomic DNA*, Pennsylvania State University.

Janzen T, Nolte AW, Traulsen A (2018) The breakdown of genomic ancestry blocks in hybrid lineages given a finite number of recombination sites. *Evolution* **72**, 735-750.

Jones FC, Grabherr MG, Chan YF*, et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55-61.

Kawakami T, Smeds L, Backström N*, et al.* (2014) A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Molecular Ecology* **23**, 4035-4058.

Kinsella CM, Ruiz-Ruano FJ, Dion-Côté A-M*, et al.* (2019) Programmed DNA elimination of germline development genes in songbirds. *Nature Communications* **10**, 5468.

Klingström T, Bongcam-Rudloff E, Pettersson OV (2018) A comprehensive model of DNA fragmentation for the preservation of High Molecular Weight DNA. *bioRxiv*, 254276.

Kupper C, Stocks M, Risse JE*, et al.* (2016) A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Genetics* **48**, 79-83.

Lamichhaney S, Berglund J, Almen MS*, et al.* (2015) Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**, 371-375.

Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of Population Structure using Dense Haplotype Data. *PLoS Genet* **8**, e1002453.

Leitwein M, Duranton M, Rougemont Q, Gagnaire P-A, Bernatchez L (2020) Using Haplotype Information for Conservation Genomics. *Trends in Ecology & Evolution* **35**, 245-258.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**, 1754-1760.

Loh P-R, Palamara PF, Price AL (2016) Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics* **48**, 811-816.

Martin SH, Dasmahapatra KK, Nadeau NJ*, et al.* (2013) Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Research* **23**, 1817-1828.

Martin SH, Jiggins CD (2017) Interpreting the genomic landscape of introgression. *Curr Opin Genet Dev* **47**, 69-74.

McKenna A, Hanna M, Banks E*, et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303.

Meier JI, Marques DA, Mwaiko S*, et al.* (2017) Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications* **8**, 14363.

Moorjani P, Sankararaman S, Fu Q*, et al.* (2016) A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. *Proceedings of the National Academy of Sciences* **113**, 5652-5657.

Nadachowska-Brzyska K, Burri R, Ellegren H (2019) Footprints of adaptive evolution revealed by whole Z chromosomes haplotypes in flycatchers. *Molecular Ecology* **28**, 2290–2304.

Nurk S, Walenz BP, Rhie A*, et al.* (2020) HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *bioRxiv*, 2020.2003.2014.992248.

O'Connell J, Sharp K, Shrine N*, et al.* (2016) Haplotype estimation for biobank-scale data sets. *Nature Genetics* **48**, 817-820.

Palamara PF, Lencz T, Darvasi A, Pe'er I (2012) Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History. *The American Journal of Human Genetics* **91**, 809-822.

Palamara PF, Pe'er I (2013) Inference of historical migration rates via haplotype sharing. *Bioinformatics* **29**, i180-i188.

Pool JE, Nielsen R (2009) Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* **181**, 711-719.

Porubsky D, Garg S, Sanders AD*, et al.* (2017) Dense and accurate whole-chromosome haplotyping of individual genomes. *Nature Communications* **8**, 1293.

R Core Team (2020) *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ravinet M, Faria R, Butlin RK*, et al.* (2017) Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology* **30**, 1450-1477.

Redin D, Frick T, Aghelpasand H*, et al.* (2019) High throughput barcoding method for genome-scale phasing. *Scientific Reports* **9**, 18116.

Sabeti PC, Reich DE, Higgins JM*, et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832-837.

Santure AW, Garant D (2018) Wild GWAS—association mapping in natural populations. *Molecular Ecology Resources* **18**, 729-738.

Schielzeth H, Rios Villamil A, Burri R (2018) Success and failure in replication of genotype-phenotype associations: How does replication help in understanding the genetic basis of phenotypic variation in outbred populations? *Molecular Ecology Resources* **4**, 739-754.

Schweizer M, Warmuth V, Alaei Kakhki N*, et al.* (2019a) Parallel plumage color evolution and pervasive hybridization in wheatears. *Journal of Evolutionary Biology* **32**, 100-110.

Schweizer M, Warmuth VM, Alaei Kakhki N*, et al.* (2019b) Genome-wide evidence supports mitochondrial relationships and pervasive parallel phenotypic evolution in open-habitat chats. *Molecular Phylogenetics and Evolution* **139**, 106568.

Sedghifar A, Brandvain Y, Ralph P (2016) Beyond clines: lineages and haplotype blocks in hybrid zones. *Molecular Ecology* **25**, 2559-2576.

Shchur V, Svedberg J, Medina P, Corbett-Detig R, Nielsen R (2019) On the distribution of tract lengths during adaptive introgression. *bioRxiv*, 724815.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212.

Snyder MW, Adey A, Kitzman JO, Shendure J (2015) Haplotype-resolved genome sequencing: experimental methods and applications. *Nature Reviews Genetics* **16**, 344-358.

Soria-Carrasco V, Gompert Z, Comeault AA*, et al.* (2014) Stick Insect Genomes Reveal Natural Selection's Role in Parallel Speciation. *Science* **344**, 738-742.

Stankowski S, Chase MA, Fuiten AM*, et al.* (2019) Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. *PLOS Biology* **17**, e3000391.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**, 978-989.

Tang K, Thornton KR, Stoneking M (2007) A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. *PLOS Biology* **5**, e171.

Taylor SA, Larson EL (2019) Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nature Ecology & Evolution* **3**, 170-177.

Toomey MB, Marques CI, Andrade P*, et al.* (2018) A non-coding region near Follistatin controls head colour polymorphism in the Gouldian finch. *Proc Biol Sci* **285**.

Vijay N, Bossu CM, Poelstra JW*, et al.* (2016) Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. *Nature Communications* **7**, 13195.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A Map of Recent Positive Selection in the Human Genome. *PLOS Biology* **4**, e72.

vonHoldt BM, Pollinger JP, Earl DA*, et al.* (2011) A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome Research* **21**, 1294-1305.

Wang O, Chin R, Cheng X*, et al.* (2019) Efficient and unique co-barcoding of second-generation sequencing reads from long DNA molecules enabling cost effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Research* **29**, 798-808.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB (2017) Direct determination of diploid genome sequences. *Genome Research* **27**, 757-767.

Weissensteiner MH, Bunikis I, Catalán A*, et al.* (2019) The population genomics of structural variation in a songbird genus. *bioRxiv*, 830356.

Zheng GX, Lau BT, Schnall-Levin M*, et al.* (2016) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology* **34**, 303-311.

Data Accessibility

Short-read sequence data are available in the European Nucleotide Archive (ENA), Accession Number PRJEB38232. The Cyprus wheatear draft reference genome assembly is available on Dryad (https://doi.org/10.5061/dryad.9zw3r22bf).

Author contributions

RB designed the research with input from DL and HSch. DL, RR, RB, and HSch performed data analysis. JG and PE performed library preparation and genome sequencing. RAO performed genome assembly. MS, AS, JTG, and HShi contributed materials. RB wrote the paper with input from all authors.

Tables

Table 1 | Sample information

Individual	Species	(Museum) No. †	Year	Tissue	Coverage [x]	Mol. Length [kb] ‡
7359_104	*O. melanoleuca*	YPM 101348	2005	Muscle	68	10.6
8854_101	*O. pleschanka*	MCZ 349924	2012	Muscle	54	17.9
8854_102	*O. melanoleuca*	A1167	2000	Blood	54	47.8
8854_103	*O. hispanica*	2917	2009	Blood	52	33.1
8854_104	*O. hispanica*	2919	2009	Blood	56	58.9
8854_105	*O. pleschanka*	16	2012	Blood	46	37.3
8854_106	*O. cypriaca*	19	2012	Blood	60	63.8

†Individual identifiers and where applicable museum numbers are provide. MCZ, Museum of Comparative Zoology Harvard; YPM, Yale Peabody Museum.

‡ Mean input DNA molecule length as determined by Supernova 2.1 are provided.

Table 2 | Effect of coverage and molecule length on phase block contiguity (N50, multiple R^2=0.91, p<0.001; N90, multiple R^2=0.88, p<0.001; 206 degrees of freedom).

	N50			N90		
	Estimate	t	p	Estimate	t	p
Intercept	13.482	427.234	$< 10^{-3}$	11.907	363.976	$< 10^{-3}$
Coverage	0.181	40.725	$< 10^{-3}$	0.144	31.288	$< 10^{-3}$
Molecule Length	0.036	21.014	$< 10^{-3}$	0.037	21.041	$< 10^{-3}$
Coverage x Molecule Length	-0.001	-5.771	$< 10^{-3}$	-0.001	-5.947	$< 10^{-3}$

Table 3 | Effect of coverage, molecule length, and filtering on coverage of genome by phase sets (multiple R^2=0.66, p<0.001, 204 degrees of freedom).

	Estimate	t	p
Intercept	74.586	82.042	$< 10^{-3}$
Coverage	1.867	14.603	$< 10^{-3}$
Molecule Length	0.246	4.994	$< 10^{-3}$

Filter	-1.968	-8.372	$< 10^{-3}$
Coverage x Molecule Length	-0.028	-4.100	$< 10^{-3}$
Coverage x Filter	0.278	8.422	$< 10^{-3}$

Table 4 | Effect of coverage, filtering, and molecule length on phasing accuracy (multiple R^2=0.80, p<0.001, 204 degrees of freedom) and genotyping accuracy (multiple R^2=0.85, p<0.001, 204 degrees of freedom).

	Phasing Accuracy			Genotyping Accuracy		
	Estimate	T	p	Estimate	T	p
Intercept	-0.739	-14.785	$< 10^{-3}$	-0.0743	-21.378	$< 10^{-3}$
Coverage	0.197	28.003	0.004	1.107	31.775	$< 10^{-3}$
Filtering	0.037	2.891	$< 10^{-3}$	0. 284	8.168	$< 10^{-3}$
Molecule Length	0.009	3.348	$< 10^{-3}$	0.006	0.178	0.86
Coverage x Filtering	-0.007	-3.879	$< 10^{-3}$	0.189	5.412	$< 10^{-3}$
Coverage x Molecule Length	0.001	3.219	0.001	0.099	2.844	0.005

Figures

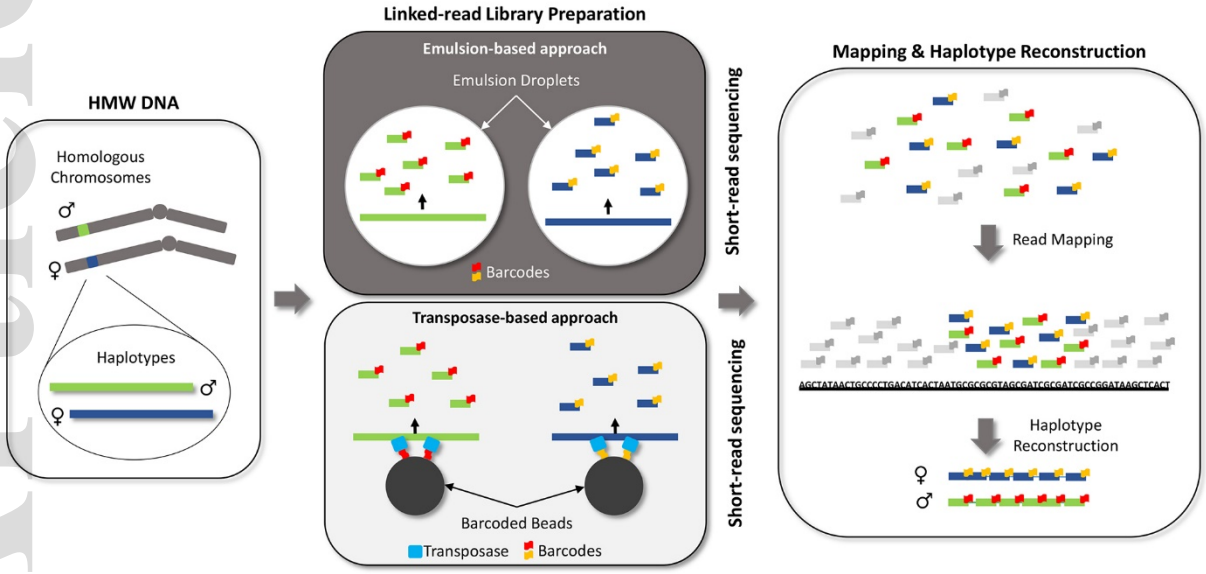

Figure 1 | Cornerstone principles of linked-read sequencing and haplotype reconstruction. After isolation of high-molecular-weight DNA, linked-read sequencing library preparation separates library preparation between DNA molecules either by introducing a limited number of DNA molecules into emulsion droplets ('emulsion-based approach') (Zheng *et al.* 2016) or by attaching DNA molecules to barcoded beads ('transposase-based approach') (Wang *et al.* 2019; Chen *et al.* 2019). In emulsion droplets, droplet-specific barcodes are ligated to DNA fragments. In the transposase-based approach, transposases insert barcodes and sequencing adapters to DNA fragments (only the transposase inserting one adapter and the barcode is depicted). The result in both cases is that DNA molecules contained in one compartment or bound by one bead carry unique barcodes. As homologous maternal and paternal chromosome stretches are unlikely to end up in/on the same compartment/bead, they carry different barcodes. After read mapping, barcodes are then used to reconstruct haplotypes from reads with the same barcode that map to the same genomic region. The schematic of library preparation shows simplifications; for detailed descriptions we refer to the above references and library preparation kit providers.

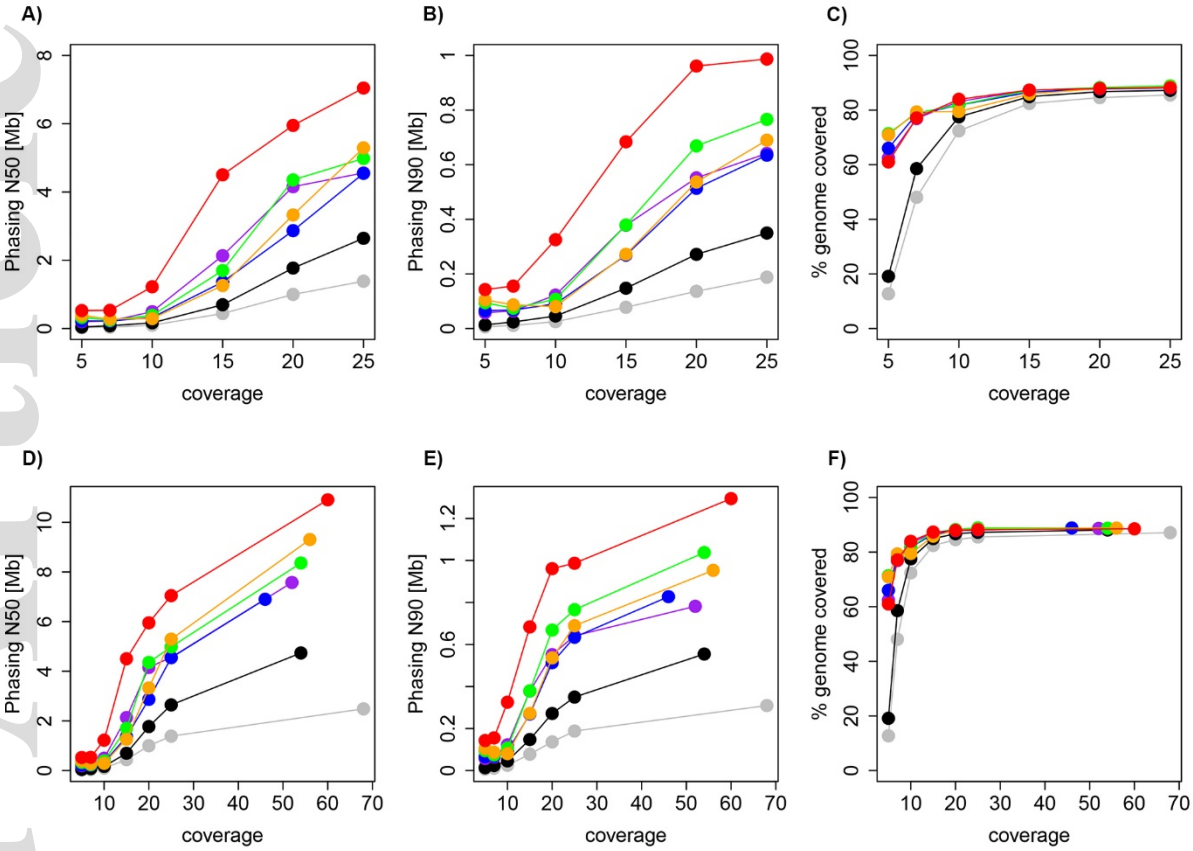

Figure 2 | Phase block contiguity (A, B, D, E) and proportion of the genome (C, F) covered by phase sets at different sequencing coverages. A-C show results for downsampled data only (for better visibility at these coverages); D-F include the full data. Colors show values for different individuals/molecule lengths. Molecule lengths (kb): Red, 63.9; orange, 58.9; green, 47.8; blue, 37.3; purple, 33.1; black, 17.9; grey, 10.6

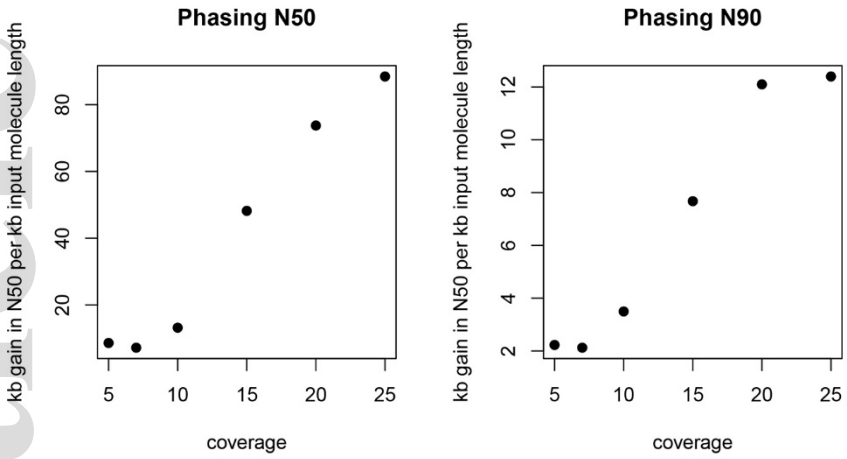

Figure 3 | Gain in phase block contiguity through increased input molecule lengths at different coverages.

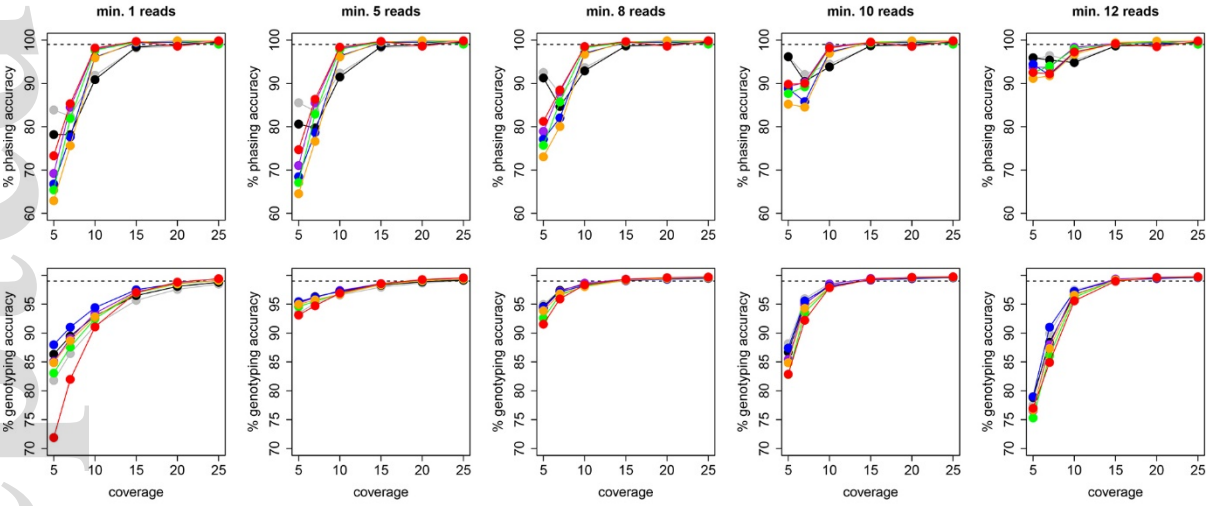

Figure 4 | Phasing and genotyping accuracies at different coverages and read-depth filtering. Different colors show values for different individuals/molecule lengths. Molecule lengths (kb): Red, 63.9; orange, 58.9; green, 47.8; blue, 37.3; purple, 33.1; black, 17.9; grey, 10.6.