APPLICATIONS OF NEXT-GENERATION SEQUENCING

Haplotype-resolved genome sequencing: experimental methods and applications

Matthew W. Snyder[1], Andrew Adey[2], Jacob O. Kitzman[3,4] and Jay Shendure[1]

Abstract | Human genomes are diploid and, for their complete description and interpretation, it is necessary not only to discover the variation they contain but also to arrange it onto chromosomal haplotypes. Although whole-genome sequencing is becoming increasingly routine, nearly all such individual genomes are mostly unresolved with respect to haplotype, particularly for rare alleles, which remain poorly resolved by inferential methods. Here, we review emerging technologies for experimentally resolving (that is, 'phasing') haplotypes across individual whole-genome sequences. We also discuss computational methods relevant to their implementation, metrics for assessing their accuracy and completeness, and the relevance of haplotype information to applications of genome sequencing in research and clinical medicine.

Haplotypes — sequences of genetic variants that co-occur along single chromosomes — are an essential concept in genetics. Haplotype information has a crucial role in diverse contexts, including linkage analysis, association studies, population genetics and clinical genetics. For example, reference panels of phased haplotypes are now commonly used to improve the power of genome-wide association studies[1–4]; studies of human population history, migration patterns and bottlenecks can gain deeper insights into the past with increased precision when analysing haplotypes rather than unphased genotypes[5–7]; and the accurate haplotype assignment (that is, 'phasing') of protein-altering alleles in drug metabolism genes is important to minimize the burden of adverse drug reactions[8,9]. However, a key limitation of contemporary genome-wide genotyping technologies — such as high-density single-nucleotide polymorphism (SNP) microarrays or whole-genome sequencing (WGS) using next-generation sequencing (NGS) platforms — is that they provide little, if any, haplotype information at the level of an individual genome.

Methods for resolving haplotypes can be broadly divided into inferential methods, which statistically infer haplotypes from the unphased genotypes of multiple related or unrelated individuals, and direct methods, which apply specialized experimental techniques to genomic DNA derived from a single individual. As others have recently reviewed both population-based and pedigree-based inferential methods[10], we restrict our focus here to direct methods as well as to the combination of direct and inferential methods. However, despite the modest cost and high scalability of inferential methods, we emphasize at the outset that it is the limitations of these methods that motivated the development of direct methods. Population-based haplotype inference, which is based on the genotyping of multiple unrelated individuals, is challenged by low-frequency variants, private variants and *de novo* variants, and is limited by the magnitude and extent of linkage disequilibrium, which differs depending on ancestry and decays with increasing genomic distance[11]. Pedigree-based haplotype inference requires the genotyping of multiple individuals from the same family; therefore, it depends on the availability of such samples and is unable to phase *de novo* variation in the last generation. By contrast, direct methods have the potential to fully resolve haplotypes for all forms of variation genome-wide using only the sample of interest. Moreover, current limitations of direct methods can be partially overcome through their combined application with inferential methods.

Of note, the first two assembled human genomes contained extensive haplotype information, at least at the local scale. The Human Genome Project was primarily executed through the hierarchical sequencing of large-insert clones — that is, 50–200-kb bacterial

[1]*Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA.*
[2]*Department of Molecular and Medical Genetics, Oregon Health & Sciences University, Portland, Oregon 97239, USA.*
[3]*Department of Human Genetics, University of Michigan, Ann Arbor, Michigan 48109, USA.*
[4]*Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, USA.*
Correspondence to M.W.S. and J.S.
e-mails: mwsnyder@uw.edu; shendure@uw.edu

doi:10.1038/nrg3903
Published online 7 May 2015

Box 1 | Targeted haplotyping

Both dense and sparse direct methods for haplotype determination typically yield genome-wide haplotype assemblies. In numerous contexts — for example, when assaying the phase of recessive mutations in disease-associated genes — obtaining local or targeted haplotypes may be desirable. Although genome-wide assemblies can be masked or partitioned to retrieve only desired targets, several targeted haplotyping techniques have been developed to reduce the complexity and/or cost relative to genome-wide direct methods. A representative subset of these is reviewed below.

One methodology for local haplotype determination involves the imaging of DNA fragments on a glass surface. As described by Xiao *et al.*[82], a small number of nearby variants are targeted with padlock probes in the presence of fluorescently labelled nucleotides[81] or with allele-specific primers with fluorescent 5′ ends. By first stretching these fragments on the surface of a microscope slide and then visualizing the fluorescent signals of and distances between adjacent probes on the stretched fragments, haplotypes of up to tens of kilobases and containing a handful of variants can be determined. By contrast, targeted (albeit sparse) haplotypes can be obtained by polony haplotyping[42,83], in which limited amounts of high-molecular-weight (HMW) DNA or entire chromosomes are immobilized in a polyacrylamide gel matrix on a glass surface, amplified and subjected to multiple rounds of fluorescently labelled single-base extension using primers that are specific to a locus of interest.

By compartmentalizing fragments of DNA along with necessary reagents in emulsions, targeted haplotype interrogation can take place simultaneously inside many individual reaction chambers. In one early application, Wetmur *et al.*[84] ascertained the phase of three heterozygous sites in paraoxonase 1 (*PON1*) by performing linking PCR using biotinylated primers within each droplet, followed by allele-specific PCR for haplotype determination. Turner and colleagues[85] developed an emulsion PCR-based method called haplotype fusion to study inversion breakpoints, performing single-molecule amplification on beads in emulsions after a fusion PCR step to juxtapose the sequences adjacent to inversion breakpoints. More recently, Regan and co-workers[86] described Drop-Phase, a scalable and rapid method based on digital droplet PCR that enables the determination of local haplotypes at a small number of sites. The maximum length of the haplotype in this method is primarily a function of the length of the input DNA template molecules.

If a library of bacterial artificial chromosomes (BACs) or fosmids is available, targeted haplotypes can be ‘dialled out’ from this clone pool. Early brute-force methods involved plating and manually screening thousands of clones. Nedelkova *et al.*[87] combined PCR-based screening of diluted pools with a recombineering strategy to extract and sequence specific ~35-kb haplotypes. Although this approach can produce dense accurate haplotypes, the experimental complexity rivals that of many genome-wide direct methods.

More promising are recent advances that are amenable to automation and high-throughput sequencing. Kaper *et al.*[32] coupled a dilution-based haplotyping approach with an additional target enrichment step, using capture baits to pull down a locus of interest from sub-haploid sequencing libraries. The resulting 1-Mb haplotypes densely phased nearly all of the ~1,200 variant sites in the dystrophin (*DMD*) gene. Despite the high accuracy and comprehensiveness of this method, it remains challenging because of the need to design and synthesize libraries of baits specific to each locus. More recently, de Vree *et al.*[88] adapted the framework of chromatin interaction maps, previously used to generate sparse genome-wide haplotypes[55], to a targeted approach termed targeted locus amplification (TLA). By using locus-specific primers, libraries derived from the crosslinking and ligation of physically adjacent regions *in vivo* are enriched for alleles falling within several hundred kilobases of a genomic target. Although still limited by the input requirement for large numbers of cells and thus to samples of tissue or cell lines, TLA enables the targeting of multiple loci per sample (by increasing the number of primer pairs) and multiplexing to large numbers of samples (by automation and indexing).

artificial chromosomes (BACs) — each of which represented a single haplotype[12]. As most of the BAC clones sequenced by the Human Genome Project were derived from a single African-American individual, the human reference genome consists of a ‘patchwork’ of experimentally phased BAC-length haplotypes of African or European ancestry[13,14]. The first personal genome — known as HuRef — was also experimentally phased with key contributions from computational haplotype assembly methods, exploiting mate-paired Sanger sequencing reads from diverse cloning strategies to resolve most of its variation into local haplotype blocks, some of which were long enough to span typical genes (N50 of 350 kb)[15,16]. N50 is the current standard descriptor of the contiguity achieved in experimental haplotyping and is defined, in the context of haplotype assemblies, as the smallest haplotype block in which the sum of that block and all larger blocks total to 50% (by length) of the complete haplotype assembly. Thus, an N50 of 350 kb means that 50% of the haplotype-resolved sequence is within blocks of at least 350 kb.

Since 2005, NGS technologies have enabled a >100,000-fold reduction in the cost of DNA sequencing, such that many thousands of individual human genome sequences are being generated — for example, by the 1000 Genomes Project[11,17]. However, the fairly short read lengths of the most cost-effective platforms — currently up to 150 bp in paired-end sequencing — make it challenging to exploit these data to extensively resolve haplotypes, as most reads span no more than a single variant. In principle, this could be overcome through diverse mate-pair distances, as was done for HuRef[12,15,16]. However, as standard mate-paired NGS library preparation protocols require an *in vitro* circularization step, sufficiently complex mate-paired libraries are constrained to a few kilobases, and the resulting haplotype reconstructions are poor. For example, in one study, fewer than half of the single-nucleotide

Low-frequency variants
Single-nucleotide variants, insertions and deletions (indels) or copy-number variants that have minor allele frequency in a population < 1 %; that is, variants found on < 1 out of every 100 haplotypes.

Private variants
Variants that are found in a single individual or pedigree and are thus recalcitrant to phasing by population-based methods owing to their absence from reference panels.

Linkage disequilibrium
A measure of the probability that two polymorphic loci do not segregate independently within a population.

Padlock probes
Single-stranded DNA oligonucleotides that have a constant region flanked by two targeting ‘arms’ that are complementary to the sequence of a genomic target. After highly specific hybridization to the target, the probes can be circularized and analysed for genotyping.

Personal genome
A substantially complete genome sequence of a single individual, typically obtained to attempt to describe or predict medical or other traits of that individual.

Mate-paired
Pertaining to a type of sequencing library preparation in which portions of a haplotype separated by 3–5 kb are brought into proximity by fragmentation and *in vitro* circularization. By sequencing across the junction of these circles, variants that are separated in genomic space but that appear on the same fragments can be jointly phased.

variants (SNVs) were phased, and nearly all were in blocks no greater than the maximum insert size of 3.5 kb[18]. As a result, almost all of the many thousands of human genome sequences that have been generated to date were phased solely through population- or pedigree-based inference.

Over the past few years, a number of direct methods have been developed to enable NGS-based haplotype-resolved WGS. Here, we review these experimental strategies, focusing on their relative advantages and disadvantages. Additional topics that we discuss include computational methods for resolving haplotypes from these experimental data sets, metrics for assessing the accuracy and completeness of genome-wide haplotype resolution, the integration of direct and inferential methods, and applications of haplotype-resolved human genome sequencing in biomedical research and clinical medicine.

Experimental methods

In the nomenclature used here, direct methods resolve haplotypes through the experimental analysis of an individual sample, which generally consists of either high-molecular-weight (HMW) genomic DNA, or intact cells or tissue, and we use the terms resolve and phase interchangeably. This section focuses primarily on experimental methods for genome-wide haplotyping, although selected methods for targeted haplotyping are briefly reviewed as well (BOX 1).

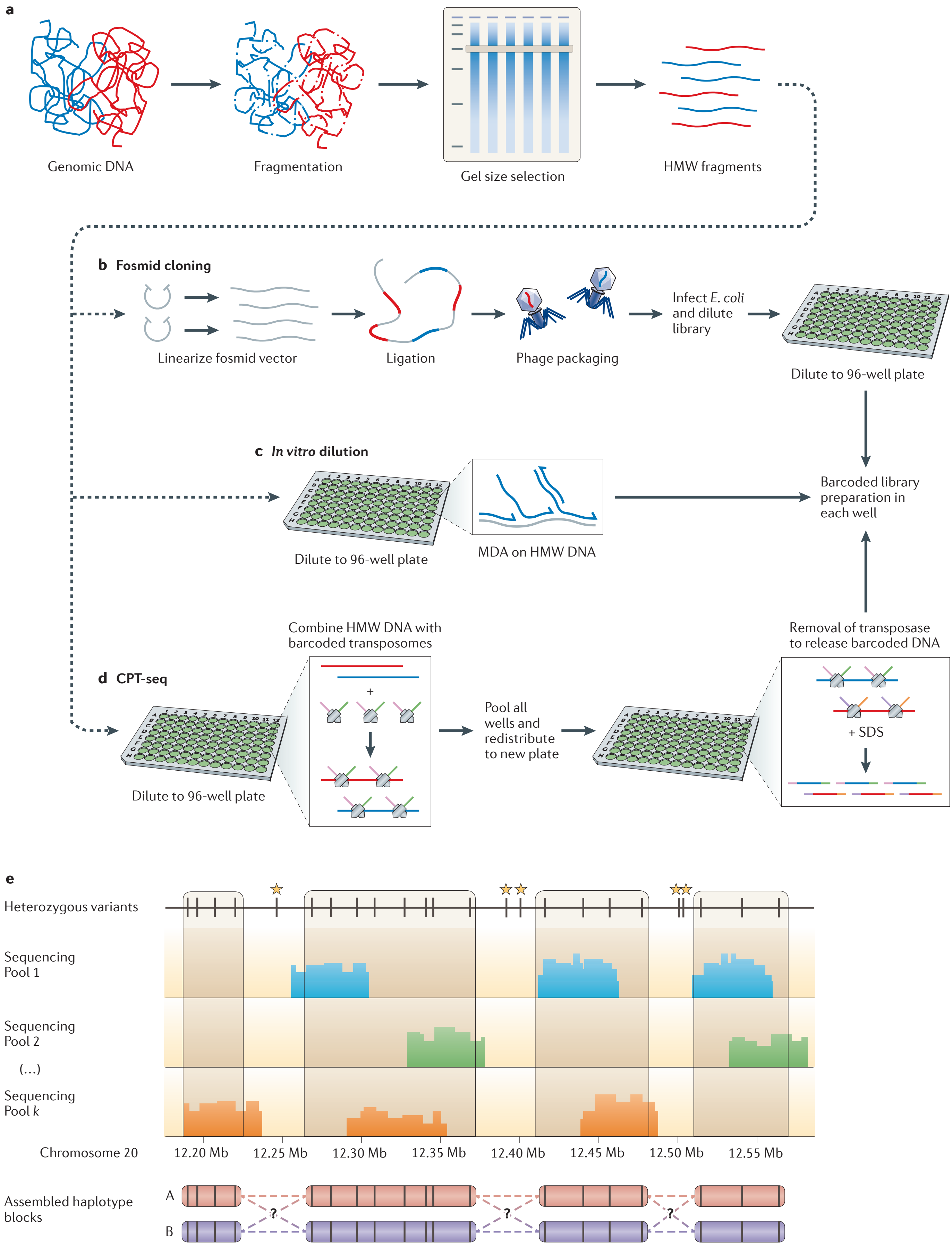
Based on the characteristics of the information that they ultimately provide, direct methods for genome-wide haplotyping broadly fall into two classes, which are referred to here as dense and sparse methods. Dense direct methods extensively resolve local haplotypes, so that any given heterozygous variant is successfully phased with respect to the other variants in the same region, yielding haplotype blocks that are typically hundreds of kilobases to several megabases in length. However, little or no experimental information relates these haplotype blocks to other such blocks on the same chromosome. Sparse direct methods leave many individual variants unphased but provide phase information for a subset of variants across much longer physical distances — up to entire chromosomes.

With a few exceptions, in the methods described here, genotyping is performed separately from haplotyping. In other words, shotgun WGS is used to generate a catalogue of unphased heterozygous variants, which are then phased by additional experimental work and sequencing.

Dense methods for genome-wide experimental haplotyping. Dense methods for genome-wide haplotyping all rely on a shared principle — pioneered by Dear and Cook[19] and refined in key ways by Sauer[20] and Olson[21] — which is to compartmentalize pools of HMW genomic DNA fragments by limiting dilution such that, within each pool, genomic regions are overwhelmingly represented either only once or not at all (FIG. 1). Each pool is thus sub-haploid in its content. However, the pools collectively include redundant coverage of the entire genome. For example, 100 pools, each of which contains a random sampling of ~3,000 HMW genomic DNA fragments of ~50 kb that sum to 5% of the haploid genome length, collectively provide ~5× coverage. Each pool is then converted to a shotgun sequencing library and indexed with DNA barcode tags so that many tagged libraries can be combined for massively parallel sequencing. Within each pool, sequence reads mapping to a given genomic region — typically appearing as an 'island' of coverage — overwhelmingly originate from a single HMW genomic DNA fragment (that is, a single haplotype). Therefore, within a given island, the alleles observed for heterozygous variants can be assigned to the same haplotype. The size of the

Figure 1 | **Schematic of dense haplotyping methods.** **a** | Genomic DNA is gently extracted from a population of cells. The resulting sample contains a mixture of both haplotypes (blue and red fragments) from every genomic region. After gentle fragmentation, high-molecular-weight (HMW) DNA may be selected by size, in this example, by gel electrophoresis, to enrich for fragments of tens to hundreds of kilobases in length. **b** | The resulting pool of HMW fragments may be cloned into a fosmid vector, packaged in phage and used to transduce *Escherichia coli* for library propagation and outgrowth. The resulting library is randomly diluted to a large number of reaction chambers (for example, $n = 96$), such that each chamber has zero or one copy of any given genomic region and thus no more than a single haplotype at any locus. An indexed sequencing library is prepared separately from each diluted pool. **c** | Alternatively, the HMW fragments may be randomly diluted to a large number of reaction chambers (for example, $n = 96$) so that each chamber has a sub-haploid genome representation. After whole-genome amplification by multiple displacement amplification (MDA), an indexed sequencing library is prepared from each amplified pool. **d** | In contiguity-preserving transposition sequencing (CPT-seq), the HMW fragments are randomly diluted into a large number of reaction chambers (for example, $n = 96$) and combined with barcoded transposomes, causing each fragment to be tagged many times while preserving structural contiguity. After transposition, the fragments from all of the chambers are pooled and again diluted to another set of reaction chambers. A protein denaturation step completes the fragmentation of the HMW DNA, and a second set of barcodes and sequencing adaptors are added by PCR in each well. **e** | In each method, all libraries are sequenced, and reads from each pool are separately aligned to the genome and compared to a list of heterozygous sites produced by an orthogonal method (for example, conventional shotgun sequencing followed by variant calling). 'Islands' of coverage overlapping between two or more pools (shaded boxes) are stitched together on the basis of allele sharing at heterozygous sites. The resulting haplotype blocks (A and B) densely cover the variation within the windows defined by overlapping islands of coverage, and few variants are left unphased (yellow stars). However, contiguity between adjacent blocks is unknown, and haplotypes longer than one to two megabases are typically not obtainable. Part **d** from REF. 34, Nature Publishing Group.

High-molecular-weight (HMW) genomic DNA
Genomic DNA isolated in such a way as to preserve long intact DNA fragments, ideally exceeding 100 kb on average. The ideal length may differ depending on the application.


a
Genomic DNA
Fragmentation
Gel size selection
HMW fragments
b Fosmid cloning
Linearize fosmid vector
Ligation
Phage packaging
Infect E. coli and dilute library
Dilute to 96-well plate
c In vitro dilution
Dilute to 96-well plate
MDA on HMW DNA
Barcoded library preparation in each well
d CPT-seq
Dilute to 96-well plate
Combine HMW DNA with barcoded transposomes
+
Pool all wells and redistribute to new plate
Removal of transposase to release barcoded DNA
+ SDS
e
Heterozygous variants
Sequencing Pool 1
Sequencing Pool 2
(…)
Sequencing Pool k
Chromosome 20
12.20 Mb
12.25 Mb
12.30 Mb
12.35 Mb
12.40 Mb
12.45 Mb
12.50 Mb
12.55 Mb
Assembled haplotype blocks
A
B
?
?
?

resulting haplotype blocks is therefore inherently limited by the length distribution of the HMW genomic DNA fragments. However, overlaps between heterozygous variants that are present on different fragments in different pools can be used to merge blocks, thereby increasing the contiguity and completeness of the resulting haplotype assembly.

Although dense methods for haplotype-resolved genome sequencing share a common basis, they diverge considerably in implementation, particularly in the means by which they achieve the initial compartmentalization of sub-haploid fragment pools (FIG. 1; TABLE 1). The first report of dense direct genome-wide haplotyping of an individual genome using NGS, by Kitzman *et al.*[22] in 2011, involved the cloning of genomic DNA to a single complex large-insert library, which was split into 115 pools that each contained ~3% representation of the diploid genome of an individual (~5,000 fosmids of 37 kb) (FIG. 1b). Indexed libraries corresponding to each fosmid pool were combined and sequenced, and the data were used to phase 94% of the separately ascertained heterozygous SNVs into long haplotype blocks (N50 of 386 kb). Lo *et al.*[23] recently implemented a similar approach with longer clones (~140-kb BAC inserts), achieving dense haplotype-resolved genome sequencing to an N50 of 2.6 Mb. An obvious shortcoming of this approach is that although only a single clone library is required per individual genome and sub-haploid pools are achieved by simple dilution before outgrowth, large-insert cloning is technically challenging and not readily scalable. Nonetheless, it is worth noting that, at least in the published literature, more human genomes (at least 30) have been experimentally phased by fosmid or BAC clone dilution pools than by any other method[22–29].

It may be advantageous to perform compartmentalization and amplification entirely *in vitro* — for example, by replacing large-insert cloning and *Escherichia coli*-based outgrowth for amplification with simple *in vitro* dilution and multiple displacement amplification (MDA), an approach pioneered by Paul and Apgar[30] for targeted haplotyping of the human leukocyte antigen (*HLA*) locus. In the first report of genome-wide haplotyping without cloning, Peters *et al.*[31] extended and updated this approach with DNA barcodes and NGS, in this case, on the Complete Genomics sequencing platform (FIG. 1c). They applied it to 7 human genomes, phasing 84–97% of ascertained heterozygous variants to haplotype blocks with N50 values ranging from 411 kb to 1.6 Mb[31]. More recently, Kaper *et al.*[32] demonstrated a similar approach with the Illumina sequencing platform on 2 human genomes, phasing ~95% of ascertained heterozygous variants to N50 values of 358 kb and 702 kb. Finally, Kuleshov *et al.*[33] adapted the Moleculo system to carry out *in vitro* dilution and PCR of sub-haploid pools of 7–9-kb fragments, yielding haplotype blocks of ~60 kb. The advantages of *in vitro* dilution include the avoidance of large-insert cloning (which is technically challenging and constrains the length of HMW genomic DNA fragments) and its very low input requirements (on the order of ~100 pg of genomic DNA[31]). There are also disadvantages relative to cloning, including the challenge of consistent dilution of HMW genomic DNA to each pool, non-uniform lengths of HMW genomic DNA fragments and non-uniform representation of the fragments introduced by MDA (for example, bias against GC-rich sequences). A limitation shared by cloning and *in vitro* dilution pool sequencing is the number of pools from which indexed sequencing libraries must be constructed.

To address some of these limitations, Amini *et al.*[34] recently reported a distinct *in vitro* method for dense direct haplotyping, termed contiguity-preserving transposition sequencing (CPT-seq) (FIG. 1d). In brief, CPT-seq exploits an inherent property of the Tn5 transposase — the enzyme remains tightly bound to target HMW DNA after 'tagmentation' with indexed DNA adaptors. Contiguity between alleles co-occurring on the same HMW fragment — that is, on the same haplotype — is structurally maintained in this step, while simultaneously enabling each HMW fragment to be tagged many times with the same index. After this step, HMW DNA fragments from differently indexed transposition reactions ($n = 96$) are pooled and then rediluted to assort randomly into new pools. Within this second set of pools ($n = 96$), a protein denaturation step releases the enzymatically fragmented templates, which are then amplified by PCR to introduce a second index. As this amplification step operates on ~200-bp fragments, rather than on HMW DNA, and uses constant rather than random priming, the resulting library uniformity is improved relative to other *in vitro* approaches. The use of 96 pools at each stage results in $96 \times 96 = 9{,}216$ distinct index combinations (also known as 'virtual compartments'), each of which effectively has a sub-haploid representation of the genome. With this approach, Amini *et al.*[34] were able to phase >95% of variants into blocks with N50 values of 1.4–2.3 Mb. Advantages of CPT-seq over the methods of Peters *et al.*[31] and Kaper *et al.*[32] include the large effective number of virtual compartments per physical compartment and the avoidance of MDA-associated amplification biases. A disadvantage of the current CPT-seq protocol is that only a small portion of the DNA in each virtual compartment is sequenced such that the overall amount of DNA required is higher (nanograms rather than picograms). Other technologies, such as using barcoded beads in emulsions, as recently described by a company named 10X Genomics[35], may represent alternative methods to efficiently achieve large numbers of compartments for the purpose of genome-wide haplotyping.

Lo *et al.*[23] recently undertook a general analysis of the parameters that underlie the success of dense methods for genome-wide haplotyping. Specifically, they modelled the impact of DNA fragment length, the number of pools, DNA fragment coverage and sequencing coverage. Although all of the parameters were relevant to some degree, they found DNA fragment length to be the key contiguity-limiting parameter in haplotype blocks resulting from dense direct methods (BOX 2). This may explain, for example, the differences in performance for the dense methods described above: methods involving long-range PCR products (7–9-kb fragments; haplotype

Fosmids
DNA cloning vectors containing up to 40 kb of insert, typically packaged in bulk into phage and transfected into *Escherichia coli*, in which a library can be propagated.

Multiple displacement amplification
(MDA). A method for high-gain whole-genome amplification in which a low input mass of high-molecular-weight DNA is exponentially copied by random priming with short oligonucleotides, followed by primer extension with a strand-displacing polymerase at a constant temperature. Resulting amplicons are typically several kilobases in length.

Complete Genomics sequencing platform
A form of high-throughput short-read sequencing technology and a suite of analysis tools offered as a commercial service.

Illumina sequencing platform
The most commonly used form of high-throughput short-read sequencing that offers a low cost per base.

Moleculo system
A commercial library preparation and *in silico* method for reconstructing the sequence of a 6–10-kb fragment of DNA using short-read sequencing instruments.

Table 1 | **A comparison of direct haplotyping methods**

	Dilution pools	CPT-seq	Long-read technologies*	Single-chromosome sequencing	HaploSeq	TLA	Emulsion PCR-based methods
Principle	Dilution of HMW DNA fragments to sub-haploid genome equivalents, amplification (in bacteria or *in vitro*) and shotgun sequencing[22,23,28,31,37]	Combinatorial barcoding of sub-haploid fractions of HMW DNA through transposition with barcoded Tn5 complexes followed by barcoding PCR[34]	Sequencing across long fragments to phase covered alleles	Isolation, random amplification and shotgun sequencing of individual chromosomes[43–45]	Crosslinking and proximity ligation followed by shotgun sequencing to read fragments that are spatially close in the nucleus but distant in sequence[55]	Crosslinking and proximity ligation (as for Haploseq) followed by inverse PCR from a selected 'viewpoint' to distal interacting sites[88]	Compartmentalized genotyping of individual HMW DNA fragments at multiple pre-defined alleles[84–86]
Genome wide or locus targeted	Genome wide	Genome wide	Genome wide	Genome wide	Genome wide	Locus targeted	Locus targeted
Scale of contiguity[‡]	Up to fragment length (10–100 kb)	Local (island N50: 45–90 kb)	Local (100 kb)	Chromosome length	Majority of read-pair inserts are <1 kb but with tail up to 30 Mb	Up to 300 kb	Up to fragment length (10–100 kb)
Ascertainment density[§]	Dense (>90% SNVs phased); individual fragments densely covered	Dense (90–97% SNVs phased); however, individual fragments are sparsely covered	Dense (although current per-base error rates of >10% limit confidence in phase of individual sites)	Sparse (as a result of allelic dropout during single-chromosome amplification)[51]	Sparse[‡] (23% before imputation)	Dense	Sparse; restricted to known alleles
Input material requirements	10 pg to1 µg DNA (for *in vitro* approaches, including Phi29 or PCR-based approaches); 1–10 µg genomic DNA (for clone based approaches)	100 ng DNA	Variable	Living mitotic cells	Intact chromatinized DNA (that is, DNA from fresh or frozen tissue or cultured cells)	Intact chromatinized DNA (that is, DNA from fresh or frozen tissue or cultured cells)	Varies but typically <100 ng genomic DNA
Equipment	Standard molecular biology reagents and equipment	Specialized transposomes (not commercially available)	Long-read sequencing instruments (not all are commercially available yet)	Means of isolating individual chromosomes, including flow cytometer, and equipment for laser capture microdissection and microfluidic isolation	Standard molecular biology reagents and equipment	Standard molecular biology reagents and equipment	Droplet generator and reader (for ddPCR)
Readout and downstream analysis	Short-read WGS of barcoded pool libraries (for fragment inference and assembly)	Short-read sequencing, including custom dual-barcode layout (for fragment inference and assembly)	Individual reads bearing two or more heterozygous alleles	Short-read WGS[43,44] or genotyping[45]	Chimeric paired-end short-read sequencing	Chimeric paired-end short-read sequencing	Varies; includes cytometric counts of dye-positive droplets (ddPCR) and allele-specific qPCR Ct values
Labour intensiveness	High (especially if preparing clone libraries); all require construction of barcoded shotgun libraries (for example, in 96-well sets per individual)	Moderate	Moderate	High	High	High	High if emulsions are prepared manually; requires up-front assay establishment

Table 1 (cont.) | **A comparison of direct haplotyping methods**

	Dilution pools	CPT-seq	Long-read technologies*	Single-chromosome sequencing	HaploSeq	TLA	Emulsion PCR-based methods
Cost	High; WGS and library construction reagent costs	High; WGS and library construction reagent costs (which are unknown)	Very high; current long-read platforms lack throughput of short-read sequencing	Moderate; lower sequence costs due to sparsity	High; WGS and library construction reagent costs	Moderate; lower sequencing costs due to its targeted nature	Low reagent cost per sample once assay has been established
Notable applications	Phased genome and epigenome of HeLa cells[25], non-invasive fetal genome sequencing[26] and archaic introgression[27]	*De novo* genome assembly[89]	Phased genome and epigenome of HeLa cells[25] and hydatidiform mole single-haplotype assembly[90]	–	Phased epigenomes[77]	–	Phasing motif-disrupting enhancer SNP with low fetal haemoglobin expression haplotype[91]

CPT-seq, contiguity-preserving transposition sequencing; Ct, qPCR threshold cycle; ddPCR, droplet digital PCR; HMW, high-molecular-weight; qPCR, quantitative PCR; SNP, singe-nucleotide polymorphism; SNV, single-nucleotide variant; TLA, targeted locus amplification; WGS, whole-genome sequencing. A representative sample of dense, sparse and targeted direct haplotyping technologies is presented to illustrate the spectra of contiguity and density of resulting assemblies, input requirements and labour and equipment costs that are associated with each method. *For example, Pacific Biosciences and Oxford Nanopore. ‡Haplotype assembly may be 'scaffolded' using reference haplotype panels to improve contiguity. §Imputation from reference panels may be used to predict phase for sites in strong linkage disequilibrium with directly phased alleles to improve haplotype density (potentially at the expense of accuracy).

Long-read sequencing methods
Sequencing technologies in which either raw or computationally assembled reads exceed 1 kb, such that each read has a greater probability of capturing two or more variants on a single haplotype. They are typically associated with a higher cost per base and a lower throughput than short-read technologies.

Subassembly
An *in silico* method for reconstructing the sequence of a DNA fragment that exceeds the maximum read length of the sequencing instrument. Molecules of ~500 bp are uniquely tagged, amplified, concatemerized and randomly fragmented. Short reads capturing the tag and a random portion of the original fragment can be jointly assembled to recover the full-length sequence.

Single-molecule real-time (SMRT) sequencing
A form of sequencing technology that directly interrogates individual molecules of DNA and thus does not require library amplification before sequencing.

blocks ~60 kb)[33] were outperformed by those involving fosmid clone dilution pools (~37-kb fragments; N50 of 386 kb)[22], which in turn were outperformed by methods involving either BACs (~140-kb fragments; N50 of 2.6 Mb)[23] or *in vitro* dilution of HMW genomic DNA (fragment lengths dependent on DNA isolation protocols used; N50 ranging from 358 kb to 2.3 Mb)[31,32,34]. The importance of DNA fragment length is further suggested by the fact that many of the 'breaks' in haplotype blocks resulting from dense methods correlate with stretches of the individual human genome in which there are a paucity of heterozygous variants or in which there are large repetitive elements or segmental duplications, such that haplotyping of very long DNA fragments is required to span them.

The findings of Lo *et al.*[23] in this context are also relevant for the expected performance of long-read sequencing methods for genome-wide haplotype resolution. For example, both subassembly[36] and the Moleculo system[33,37] implement post hoc reconstruction of sequencing reads that are substantially longer than the read lengths of the NGS platforms with which they are coupled. However, these virtual reads are still <10 kb, as they must undergo PCR amplification before fragmentation. Single-molecule sequencing technologies, including single-molecule real-time (SMRT) sequencing and nanopore sequencing, now offer the possibility of more naturally obtaining long sequence reads of >10 kb[38,39]. In either embodiment, the advantage of longer read lengths is the ability to jointly phase many alleles across a multikilobase stretch of DNA from a single read without the need for sub-haploid genome compartmentalization; however, it is likely that the contiguity of haplotype assemblies resulting from these methods (when unassisted by inferential methods; see below) will be limited compared to the above-described methods that exploit much longer DNA fragments.

Sparse methods for genome-wide experimental haplotyping. There are diverse sparse methods to resolve haplotypes across much longer physical spans, but the drawback is that many individual variants are missing and left unphased (TABLE 1). Most of these methods involve the compartmentalization of one or a small number of chromosomes (such that within any given compartment only one homologue of any given chromosome is present), amplification and then genotyping of the amplification products using microarrays or NGS (FIG. 2). A classic, albeit poorly scalable, example of this strategy is somatic cell hybrids and related methods[40,41], in which single copies of one or a few human chromosomes are present within a fused mouse–human cell line and can be genotyped. An early *in vitro* implementation of this strategy, achieved by Zhang *et al.*[42] in 2006, involved *in situ* immobilization of diluted chromosomes within a polyacrylamide gel, followed by targeted genotyping by serial PCR and single-base extensions; however, this approach was limited to the haplotyping of a handful of heterozygous loci. More recently, several groups have implemented protocols relying instead on MDA and more extensive genotyping. For example, Ma *et al.*[43] demonstrated sparse haplotyping by laser capture microdissection of individual chromosomes in metaphase spreads, followed by MDA and microarray-based genome-wide genotyping (FIG. 2b). Yang *et al.*[44] used fluorescence-activated sorting to place individual chromosomes into wells of a 96-well plate, followed by MDA and NGS (FIG. 2c). Fan *et al.*[45] used microfluidic devices to separate and amplify (with MDA) individual or small pools of chromosomes from a single metaphase cell and genotyped amplification products using either microarrays or NGS (FIG. 2d). A practical limitation of all of these methods is the requirement for intact mitotic cells.

Box 2 | Guidance on genomic DNA preparation

DNA fragment length is the key parameter influencing the contiguity of a haplotype assembly, underscoring the importance of the initial preparation and handling of genomic DNA[23,86]. To achieve the requisite size distribution, it is crucial to carefully select the proper kit for DNA isolation. Column-based DNA isolation kits should be avoided because they result in excessive shearing as DNA passes through the silica membrane. Ideal protocols involve a single DNA precipitation step after lysis and protein removal steps. Compatible commercial kits include the Qiagen Gentra Puregene kit or the Agilent DNA Extraction kit, both of which have been shown to produce DNA fragments in excess of 500 kb. It is highly recommended to assay DNA fragment length by pulsed field gel electrophoresis and optimize DNA isolation parameters before initiating downstream haplotype resolution protocols. The choice of DNA isolation method should also be based on the selected haplotype resolution method. For instance, fosmid-based methods require a higher mass of isolated DNA in the ~40-kb range, whereas dilution-based MDA and transposase-based contiguity-preserving transposition sequencing (CPT-seq) methods require a much lower total mass but perform best when fragments are as long as possible.

In addition, careful sample handling after DNA isolation is essential. Pipetting steps should be performed gently to minimize damage to DNA fragments, using wide-bore pipette tips when possible. Vortex mixing of samples and multiple freeze–thaw cycles should be avoided to minimize shearing. Finally, the choice of appropriate DNA storage buffers, such as Tris-EDTA, can help to prevent the continuing degradation of high-molecular-weight fragments after isolation.

Nanopore sequencing
A method for DNA sequencing in which small changes in electrical current are detected as sequential bases of a DNA polymer pass through a 1 nm transmembrane protein or solid-state pore. As single molecules of DNA can be sequenced directly, no library amplification step is required.

Chromatin interaction maps
Sets of measurements of the pairwise 3D spatial proximity of many non-adjacent regions of genomic DNA in a nucleus, as ascertained experimentally by crosslinking chromatin, ligating together fragments of DNA that are associated with the crosslinked proteins, and sequencing.

An alternative to the physical isolation of chromosomes is to use the natural packaging of haploid complements within human gametes. For example, several groups have recently performed genome-wide haplotyping by MDA (or multiple annealing and looping-based amplification cycles (MALBAC)[46]) and genotyping of individual sperm[47–49]. Remarkably, by analysis of the haploid polar bodies, Hou *et al.*[50] also demonstrated non-destructive genome-wide haplotyping of a human oocyte. Despite the fact that the haplotypes obtained from these methods represent the products of meiotic recombination and thus differ slightly from the parental haplotypes, the small number of expected recombination events per chromosome enables long-range contiguity to be inferred or directly obtained by analysing multiple gametes in parallel. Although these methods are undoubtedly useful in specific contexts — for example, in studies of recombination rates with sperm and in clinical pre-implantation genomic screening of oocytes — the necessary tissues are not readily available in most other contexts.

In principle, the above-described methods could yield complete chromosome-scale haplotypes. Why do the sets of successfully phased genotypes tend to be incomplete? This is primarily a consequence of the non-uniformity of high-gain amplification (for example, with MDA or MALBAC) starting from limiting input (such as an isolated chromosome or a sub-haploid fragment pool)[51]. For example, in the study by Fan *et al.*[45], microarray-based genotyping of the amplified chromosomes at sites of common variation is robust, but NGS of the same amplified material phased only 46,000 heterozygous sites. This incompleteness could be rectified by deeper sequencing, by querying of larger numbers of independently amplified chromosomes, by improving uniformity of MDA and related protocols, or by combining sparse methods with dense direct methods and/or inferential methods. When extremely limited input is available — for example, a small number of cells in the context of *in vitro* fertilization — particular attention must be given to technical experimental factors, including ensuring uniform and quantitative dilutions, and preventing loss of chromosomes or large DNA fragments that may stick to the walls of the reaction chambers.

An alternative sparse approach to genome-wide haplotyping involves exploiting contact probability maps — for example, all-by-all chromatin interaction maps generated with Hi-C and related methods[52,53]. In brief, these methods subject intact cells or nuclei to protein–DNA crosslinking, before constructing sequencing libraries in which mate-paired reads capture sequences corresponding to physically interacting regions in chromatin[54]. As homologous chromosomes occupy distinct chromosomal territories, the probability of intra-homologue interactions is much higher than that of inter-homologue interactions. Selvaraj *et al.*[55] recently demonstrated this approach, termed HaploSeq, on mammalian cell lines. They phased ~95% of heterozygous variants in an F_1 mouse embryonic stem cell line (derived from a cross between *Mus musculus castaneous* and 129S4/SvJae) and ~22% of heterozygous variants in a human HapMap cell line, with the difference primarily attributable to the much lower heterozygosity of human genomes[55]. Methodological improvements that eliminate the reliance of Hi-C on restriction enzymes might be expected to improve completeness for human genome haplotyping. However, like the other sparse methods that capture chromosome-wide haplotypes, performing Hi-C requires the availability of intact cells or nuclei. This limitation might be overcome by emerging technologies for reconstituting chromatin *in vitro*[56].

Computational methods

Computational methods for haplotype resolution generally fall into two categories: haplotype phasing (that is, inferential) approaches, in which reference panels of unrelated individuals are genotyped and used to assign, probabilistically, the most likely local phase according to an underlying evolutionary model (reviewed in REF. 10); and haplotype assembly algorithms, in which local phase is assessed by identifying single reads or read pairs capturing multiple variant sites. For haplotype assembly algorithms, as discussed above, short-read NGS technologies by themselves generally provide insufficient information for effective haplotype assembly. However, when experimental data are obtained by methods such as dilution pool sequencing, reads corresponding to a single clone or a HMW genomic DNA fragment can effectively be treated as a single synthetic read by these algorithms.

Assembly methods. The computational formulation of the haplotype assembly problem is now more than a decade old[57]. Most approaches attempt to optimize an objective function, yielding inferred haplotypes that minimize one of several error criteria. One general

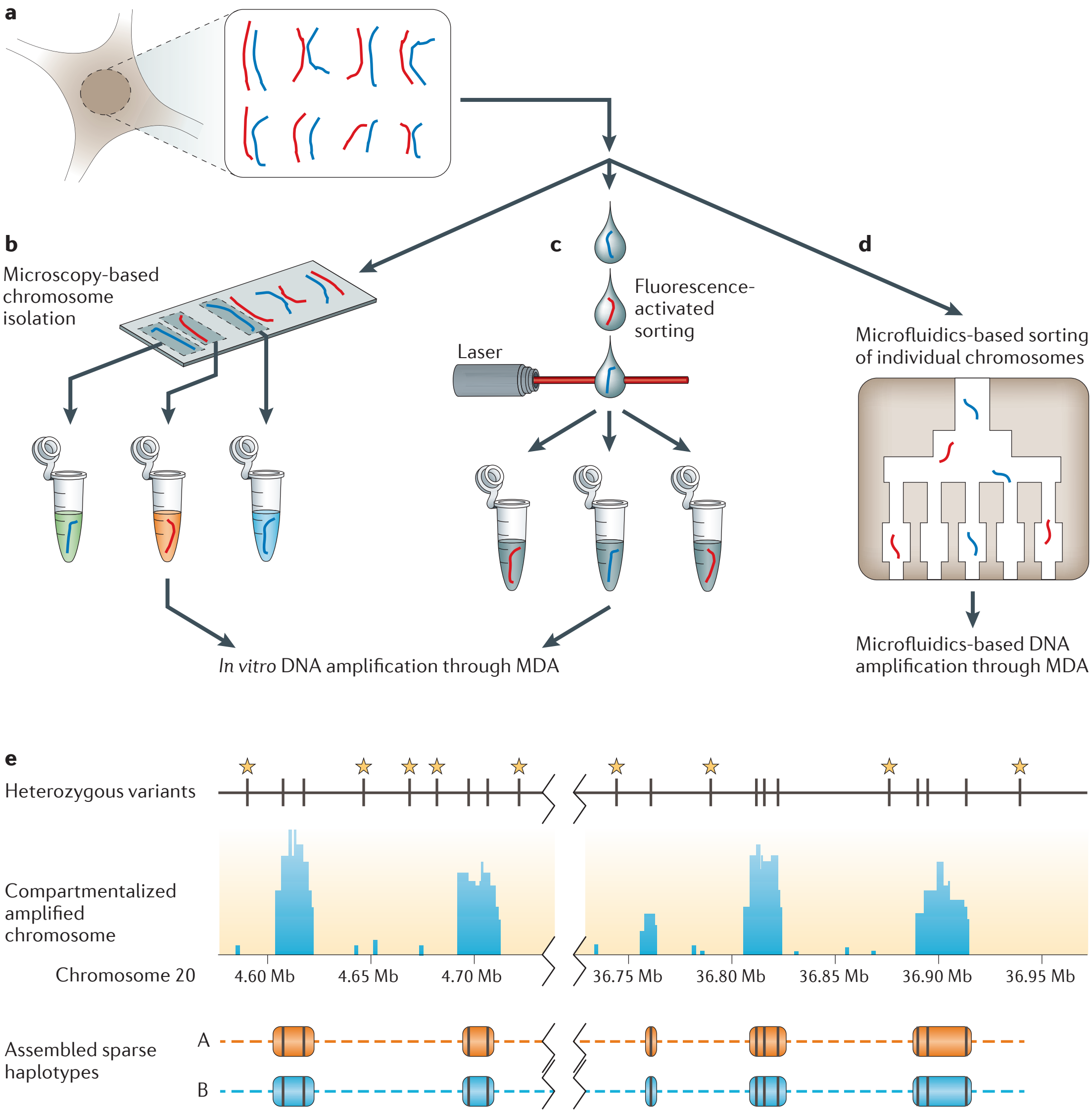

a
b
Microscopy-based chromosome isolation
c
Fluorescence-activated sorting
Laser
d
Microfluidics-based sorting of individual chromosomes
In vitro DNA amplification through MDA
Microfluidics-based DNA amplification through MDA
e
Heterozygous variants
Compartmentalized amplified chromosome
Chromosome 20
4.60 Mb
4.65 Mb
4.70 Mb
36.75 Mb
36.80 Mb
36.85 Mb
36.90 Mb
36.95 Mb
Assembled sparse haplotypes
A
B

Figure 2 | **Schematic of sparse haplotyping methods. a** | Intact metaphase chromosomes from a single nucleus are isolated and compartmentalized by one of several means. Chromosomes are sorted to separate reaction chambers, with each chamber containing no more than one chromosome and thus no more than a single haplotype. **b** | Individual chromosomes fixed to a microscope slide are microdissected and sorted into individual reaction chambers. **c** | In another method, individual chromosomes suspended in droplets are sorted by fluorescence-activated sorting into individual chambers. **d** | Alternatively, specialized microfluidic instruments are designed to lyse cells and to sort individual chromosomes into miniature reaction chambers. After compartmentalization by one of these methods, high-gain whole-genome amplification, such as multiple displacement amplification (MDA), is performed in each reaction chamber. Then, sequencing libraries are prepared from each amplified chromosome (not shown). Although uniform amplification is desirable, biases introduced during whole-genome amplification yield an uneven representation of the template in the resulting library. **e** | After sequencing, reads are aligned to the genome and compared to a list of heterozygous variants produced by an orthogonal method. As all sequenced fragments are derived from the same haplotype, heterozygous sites falling within 'islands' of coverage (blue vertical bars) can be joined to form chromosome-length haplotypes. Owing to the amplification bias during whole-genome amplification, the majority of heterozygous sites lack sequencing coverage and remain unphased (yellow stars).

strategy is to convert a set of reads or, more pertinently, a group of reads that correspond to a single clone or a HMW genomic DNA fragment into a weighted graph (the connectivity of which is determined by the density of reads covering multiple variants) and then to compute either minimum[58] or maximum[28,59] cuts on that graph.

A different graph-based approach involves applying cycle basis algorithms to solve the minimum weighted edge removal problem[60,61]. Although these approaches can achieve high accuracy over hundreds of kilobases to a few megabases, the resulting contiguity is ultimately limited by fragment length[62].

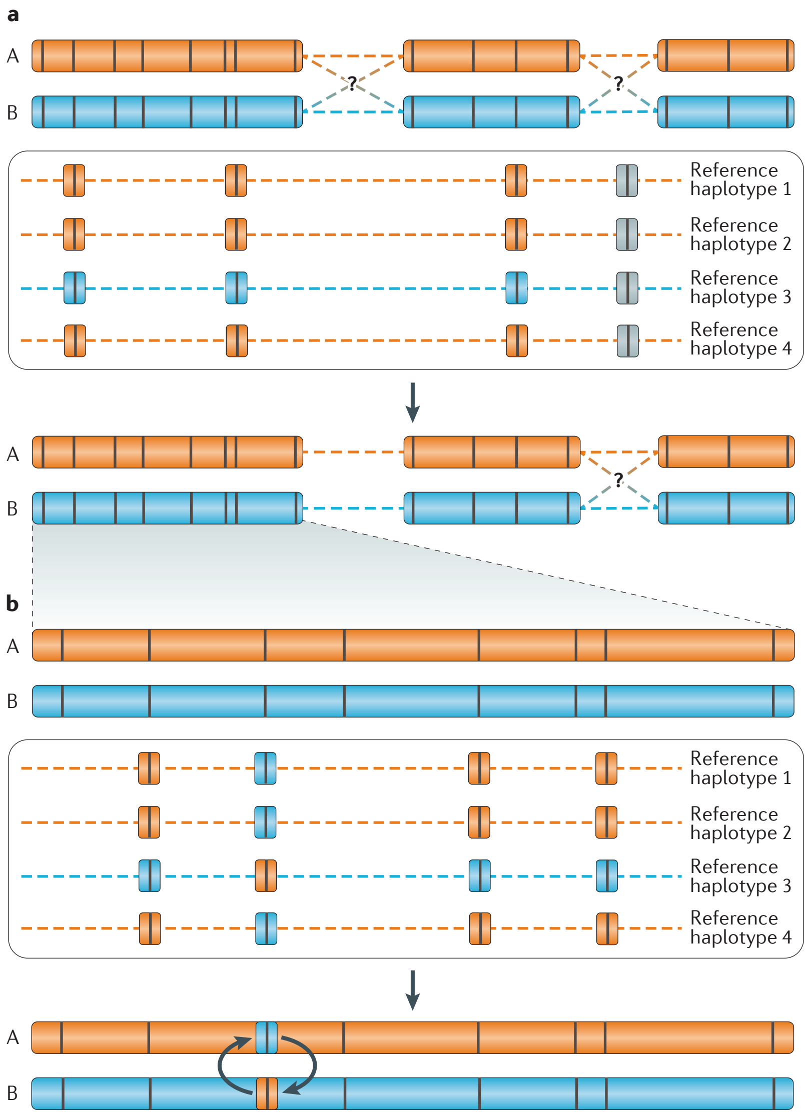


Figure 3 | **Combining direct and computational methods. a** | Dense methods for obtaining haplotypes produce phased haplotype blocks (A and B) that comprehensively encompass variation contained within the blocks. However, contiguity between adjacent blocks is unknown. Large reference panels of previously ascertained haplotypes, for example, those produced by the 1000 Genomes Project, lack many of the rare or private variants phased by direct methods but contain information about population-level linkage disequilibrium patterns between common variants. By modelling the directly phased sample as a mosaic of haplotypes segregating in the population, contiguity between pairs of nearby dense blocks can be inferred. **b** | Direct methods may produce haplotype blocks containing a small number of common variants that have experimentally determined phases which are incompatible with previously ascertained haplotypes. By again using patterns of linkage disequilibrium observed in large or population-specific reference panels, these errors can, in some cases, be corrected.

Hybrid methods. Several groups have demonstrated that a combination of population-based inference and sequencing data corresponding to the sample of interest can improve both the accuracy and the comprehensiveness of variants phased by either approach alone[63,64]. Algorithms for this include haplotyping with reference and sequencing technology (HARSH)[65], which first identifies reads or read pairs covering multiple polymorphisms and then searches phased reference panels for haplotype blocks explaining the joint presence of the alleles in these reads. FreeBayes[66] similarly uses alignments to discover polymorphic sites across multiple samples and to infer haplotypes using a Bayesian framework, notably handling multiple variant types and arbitrary ploidy at each locus. However, these algorithms rely on shotgun sequencing data rather than haplotypes resolved by the experimental methods described above.

Other groups have taken a related approach with directly resolved genome-wide haplotypes, using population-based inference to improve the end point by probabilistically assigning phase to many of the variants left unresolved by the experimental data and/or to link experimentally resolved haplotype blocks to one another[33,34,55,67]. These algorithms take advantage of the overlap between the variants successfully resolved using dense or sparse haplotyping and common polymorphisms found in increasingly large and population-specific reference panels (FIG. 3). For example, Selvaraj *et al.*[55] used patterns of linkage disequilibrium to extend sparse direct haplotypes that initially covered 22% of variants in the GM12878 genome to encompass 81% of heterozygous sites by an approach termed local conditional phasing. In this method, sparse chromosome-length haplotypes were used to seed statistical haplotype inference with BEAGLE[68] and the 1000 Genomes reference panel, resulting in whole-genome haplotypes that were fourfold denser and 98% accurate. These findings are consistent with results suggesting that, at least in well-ascertained European populations, ~80% of SNVs left unphased by direct methods can be phased with population-scale linkage disequilibrium patterns derived from reference haplotypes (M.W.S., unpublished observations). Moreover, by incorporating pairwise confidence scores arising naturally from the iterative process of statistically inferring phase, the accuracy of these predictions exceeds 97% for the highest confidence category of calls (containing ~50% of all calls), dropping to just below 90% when including all calls. An important caveat is that the haplotype information that is 'filled in' with population-based inference will be biased towards more-common variants. Nonetheless, the completeness and accuracy of this strategy will improve as reference panels are deepened and broadened to include additional populations.

On a related point, it is crucial that experimental haplotyping methods are compared before introducing information that is based on population-based inference. For example, experimental haplotyping based on the Moleculo method yields blocks that are considerably shorter than other dense methods because of the short length of the PCR-amplified fragments[33], but such haplotype blocks might be inappropriately interpreted as equivalent or superior if one compares them on the basis of performance metrics generated by a combination of experimental haplotyping and population-based inference.

Assessment of haplotype assemblies

Contiguity and accuracy. It is often desirable to summarize a haplotype assembly by a single metric in order to enable assemblies to be quickly compared to one another and to 'gold standards'. The current standard descriptor of the contiguity achieved in experimental haplotyping, particularly for dense methods, is the N50 metric. Related metrics are the S50 and AN50 metrics, which use the number of heterozygous sites or an adjusted span that accounts for interleaving between haplotype blocks, respectively[62]. Originally developed for comparative assessment of *de novo* genome assemblies, a limitation of the N50 and related metrics is that they do not take accuracy into account. By aggressively joining neighbouring haplotype blocks despite minimal experimental evidence of contiguity, the N50 of an assembly can be trivially inflated.

For genomes in which 'truth' haplotypes are known, one way to address this limitation would be to use a contiguity metric that penalizes inaccurate phasing — for example, by first breaking each haplotype block into shorter chunks at each erroneous position and then calculating the N50 of the resulting assembly. Despite the appeal of such an approach, the terms by which accuracy should be defined are not obvious and, importantly, may not be consistent for every application of haplotype-resolved sequencing.

Pairwise variant haplotype assignment accuracy approaches are attractive owing to their simplicity. The assigned phases of the alleles at any given pair of sites can be compared to gold standard calls and scored as either concordant or discordant. By performing this test for every pair of variants, binned by pairwise genomic distances, the effect on accuracy of increasing genomic distance can be quantified (FIG. 4a,b).

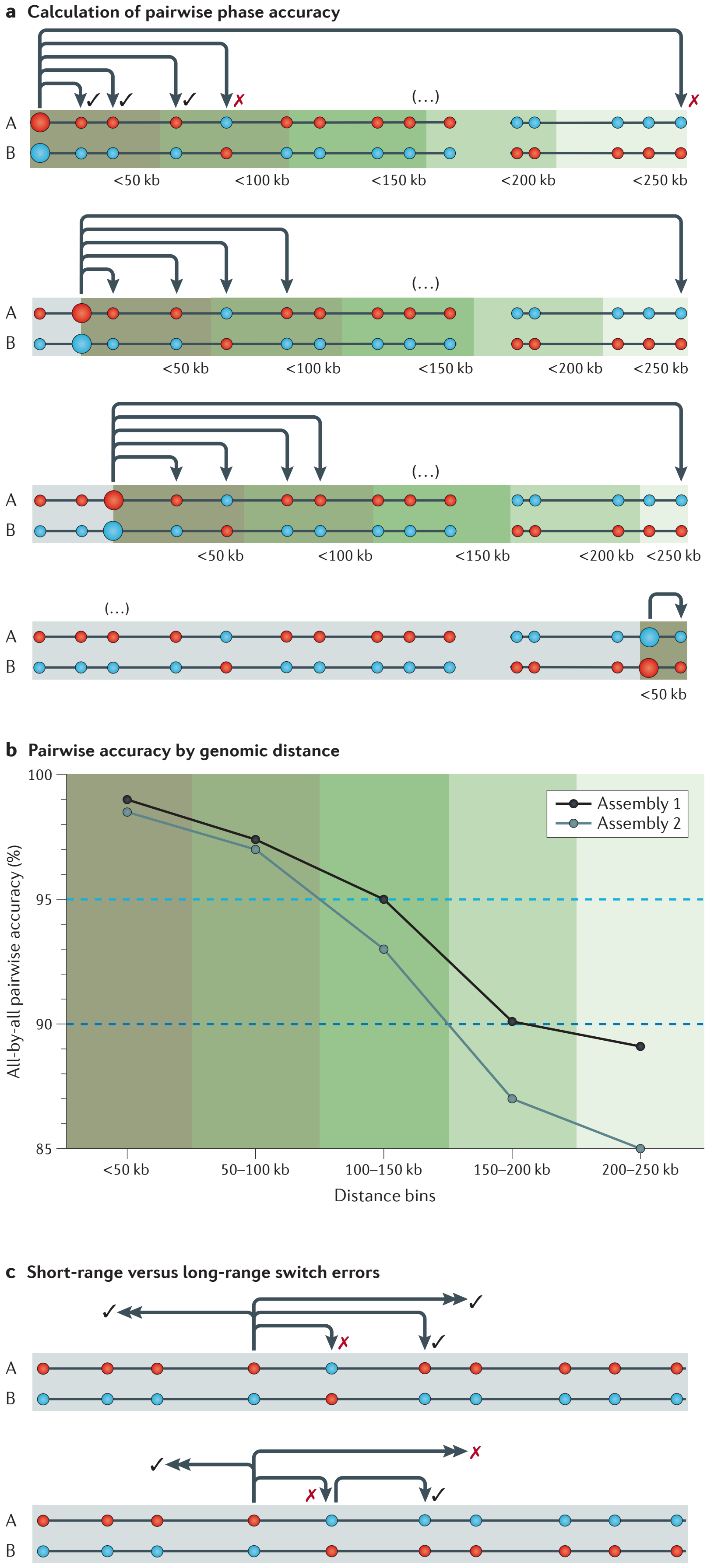
Such pairwise approaches are well-suited to capture the effect of distance on short-range switch errors, which are errors that can be fixed by flipping the phase assignment of alleles at a single site. However, they may over-penalize another class of haplotype assembly error: long-range switch errors, in which concordance with gold standard haplotypes can be attained by flipping the phase assignments at each of two or more consecutive markers (FIG. 4c). A single long-range switch error in the middle of an otherwise perfect 100-kb haplotype assembly would cause pairwise accuracy to drop to 0% in distance bins >50 kb, whereas an alternative assembly with a uniform 10% error rate would potentially compare favourably. As discussed by Kuleshov[69], comparisons between assemblies on the basis of multiple accuracy metrics may be desirable. Alternatively, the specific accuracy metric selected may depend on the downstream application of the haplotype information (discussed below).

Moreover, an overemphasis on contiguity or accuracy in isolation may not provide a complete 'snapshot' of a haplotype assembly. We propose a more comprehensive approach for assessing haplotypes based on scoring and comparing assemblies on four metrics — accuracy, contiguity, density and allele frequency spectrum. Assembly contiguity, summarized by a single N50-like statistic, provides a sense of the genomic distances over which phase has been determined. Density measures the proportion of all heterozygous variants that have been phased. Assemblies biased towards phasing common genetic variation would be scored poorly on allele frequency spectrum comparisons, whereas assemblies that are agnostic to population variant frequency would rate highly. Finally, accuracy measurements — whether expressed as rates quantifying the number of switch errors per unit of genomic distance or as pairwise accuracy over distance bins — are properly placed into context; assemblies that phase only the most confident sites may receive high accuracy marks but would score poorly on density or allele frequency metrics.

Figure 4 | **Haplotype contiguity and accuracy.** ► **a** | Haplotypes obtained by direct methods (A and B) are compared to 'gold standard' haplotypes determined by an orthogonal method (coloured circles). Beginning with the first (that is, 'index') variant in the pair of haplotypes (larger coloured circles; uppermost panel), variants are binned (green boxes) according to their genomic distance from the index variant, irrespective of gaps between adjacent haplotype blocks (horizontal black lines). Pairwise concordance between the experimentally determined and gold standard haplotypes is evaluated between the index variant and all other variants within each distance bin. Subsequently, the index variant shifts one position to the right; downstream variants are re-binned, and haplotype concordance is again evaluated within each distance bin (second panel from the top). The process iterates until the last pair of variants is assessed (bottom two panels). **b** | Concordance between all pairs of variants is aggregated within each distance bin to measure haplotype accuracy as a function of genomic distance. Specific accuracy thresholds may be used to compare multiple haplotype assemblies directly. Shown here are two fictitious assemblies; Assembly 1 retains accuracy above 95% and 90% at greater genomic distances than Assembly 2. **c** | Short-range switch errors (top panel) can be fixed by flipping the phase assignment at a single site and individually have little impact on all-by-all pairwise accuracy measurements. By contrast, long-range switch errors (bottom panel) can be fixed by flipping the phase assignments at several consecutive sites that are accurately phased relative to one another. The number of such errors per unit of genomic distance is a typical accuracy metric for inferential haplotyping methods.

Quality scores. As discussed above, many of the algorithms used in conjunction with experimental data for genome-wide haplotype resolution rely on weighted graphs. However, the extent and quality of the underlying information used to make a given call, including conflicting data, are generally not included in the output with the resulting haplotype blocks, with few exceptions[69,70]. This represents a major limitation and contrasts with genotype calls for which phred-scale quality scores are standard. The need for a quality score metric for experimental haplotypes that captures the information content and conflicting data of each phased variant and variant pair is evident; however, there are substantial challenges when attempting to implement a universal

Phred-scale quality scores
A scoring system, originally developed for assigning confidence to individual base calls from sequencing instruments, in which an estimated error probability (P) is converted to a quality score (Q) by the transformation $Q = -10\ \log_{10}(P)$.



cross-platform quality score because the algorithms used are so different. One solution may be the post hoc calibration of arbitrary quality score scales to a phred scale based on empirically determined accuracy on gold standard samples.

Comprehensiveness of variant types. Haplotypes consist of the full spectrum of genetic variation, including SNVs, short insertions and deletions (indels), structural rearrangements and copy-number polymorphisms. However, most methods for haplotype inference operate only at the level of SNVs and are restricted to unique single-copy sequences on autosomes. The development of algorithms that are capable of integrating multiple variant types into comprehensive assembled haplotypes represents an important challenge for the field.

As discussed above, many of the methods for experimental genome-wide haplotyping separate the genotyping step (data from which heterozygous variants are called) from the haplotyping step (data from which heterozygous variants are phased). Although in principle this could be extended to all forms of variation, challenges include the fact that calling of indels and structural variation from shotgun, short-read sequencing data remains challenging as well as the fact that complex structural variation may confound the algorithms used for calling haplotypes from dense or sparse experimental data. Progress towards this goal may be helped by Phase 3 of the 1000 Genomes Project, in which the phasing of all forms of genetic variation by inferential methods is an explicit goal.

Features obstructing phasing. The ability to densely phase long contiguous stretches of a chromosome is determined not only by the inherent limitations of the chosen experimental method but also by the properties of the chromosome itself. Two key features that affect haplotype contiguity are the repetitive sequence content of, and the frequency of heterozygous variants within, a particular genomic region. Runs of homozygosity that exceed the maximum bridgeable length of the chosen direct phasing method will necessarily result in a break in the haplotype assembly. For example, a 690-kb run of homozygosity — located on chromosome 15 and present in ~35% of individuals — contains the gene hexosaminidase A (*HEXA*), mutations in which have been linked to Tay–Sachs disease[71]. Fosmid clone-based haplotyping methods, which require at least two heterozygous sites approximately every 40 kb, would be unable to bridge this run of homozygosity. Similarly, long regions of repetitive sequence, including many known segmental duplications, also typically result in breaks in haplotype assemblies owing to the difficulty in discriminating between the multiple copies of highly identical sequences. By contrast, aneuploidy presents less of a challenge because of the constraint that additional chromosome copies are derived from one of the two original haplotypes and the assumption that somatic mutations are infrequent compared to germline heterozygous variants. In the case of an imbalanced haplotype copy number (for example, two copies of the maternal haplotype and one copy of the paternal haplotype), the uneven allelic ratio can be used to bridge gaps in a haplotype assembly by linking phased blocks on the basis of shared allele balance across the assembled stretches[25].

Applications

What are the applications for which genome-wide haplotype resolution is desirable or necessary? The first application is the accurate interpretation of personal genomes, particularly in the context of medical genetics. As humans are diploid organisms, haplotype information is essential to each personal genome, for instance, to assess the phase of potentially disease-causing recessive mutations (that is, compound heterozygosity)[22,29]. In pharmacogenetics, the phasing of metabolically relevant variants onto haplotypes helps to predict the drug response profiles of patients, improve dosing and reduce the extent of adverse reactions[8]. Experimental methods for haplotype resolution may also enhance the accuracy of variant calls, as haploid genotypes are easier to call than diploid genotypes[31].

Second, haplotype knowledge is useful in population genetics and human disease studies. For example, the inference of Neanderthal ancestry in non-Africans exploited the availability of a human reference genome that was derived from local segments of African and European ancestry[14]. Certain methods for the inference of historic human population sizes and bottlenecks have improved accuracy when long, dense haplotype blocks are available, owing to the ability to identify older identical-by-descent segments along analysed chromosomes[7]. More generally, haplotype inference and variant imputation are increasingly important parts of human disease studies involving large cohorts — that is, rare variant–common disease and common variant–common disease study designs. In such studies, sequencing-based discovery and direct or inferential phasing of alleles in a subset of individuals enables missing genotypes in the remaining individuals to be accurately filled in on the basis of haplotype block sharing, thereby increasing the power of association testing at low cost. The uncertainties in inferred haplotypes, particularly in populations for which reference panels are not readily available, can be mitigated or completely eliminated by haplotype-resolved genome sequencing.

Third, haplotype information can be applied to studies of biological mechanisms. One example of this is the HeLa genome, for which we and colleagues generated a haplotype-resolved genome sequence by fosmid clone dilution pool sequencing[21,22], followed by a scaffolding step to connect local haplotype blocks across full chromosome arms, using the signal of allelic bias arising from the aneuploidy of the cell line[25] (similar to an approach taken by Nik-Zainal *et al.*[72]). Genome-wide haplotype information was then used to phase epigenetic and transcriptomic data (for example, data from the Encyclopedia of DNA Elements (ENCODE) project). The majority of methods that are used to interrogate these properties rely on sequence and alignment followed by read counting to determine properties such as transcription factor binding (for example, chromatin immunoprecipitation followed by sequencing (ChIP–seq)) or gene expression levels (for example, high-throughput RNA sequencing (RNA-seq)). Assigning haplotypes to these features by the presence of phased heterozygous variants facilitated the association of epigenetic regulatory machinery and proximal gene expression in *cis*. For example, we confirmed that activation of the *MYC* oncogene in HeLa cells was specific to a single haplotype, in which the active *MYC* allele was in *cis* with the chromosomal integration site of human papillomavirus (HPV)[25]. Additionally, to explore the role of mutations in melanoma antigen family L2 (*MAGEL2*) in the aetiology of Prader–Willi syndrome in a small cohort, Schaaf *et al.*[73] phased loss-of-function variants exclusively to the unmethylated paternal haplotypes. In conjunction with methylation-dependent silencing of the maternal alleles, truncating mutations in the paternal copy yield no functional expression of *MAGEL2* and may have a pathogenic role in Prader–Willi syndrome. More generally, the phasing of epigenetic information may also be highly valuable for cataloguing and investigating mechanisms of allele-specific expression and imprinting[33,74–77].

Last, haplotype information can also facilitate non-invasive fetal genome sequencing. Accurate early inference of allelic inheritance genome-wide has the ability to simultaneously determine the risk of the thousands of individually rare, but collectively common, Mendelian disorders in a single test. In 2010, Lo *et al.*[78] reported that the entire fetal genome was represented in short cell-free DNA fragments in maternal plasma, suggesting how reconstruction of the inherited fetal genome was technically attainable. We and colleagues[26] pursued and recently reported the use of haplotype-resolved parental genomes (based on fosmid clone dilution pool sequencing[21,22]) and deep sequencing of cell-free DNA in maternal plasma (a mixture of maternal and fetal DNA) to infer the fetal genome with substantial completeness and >99% accuracy, and another group[67] achieved comparable accuracy with a similar approach. In all of these studies, generating haplotype-resolved genomes for one or both parents was essential for reconstructing the fetal genotypes and haplotypes.

It is important to note that, in many cases, dense haplotype resolution in the ~100-kb range — as opposed to chromosome-wide resolution — is sufficient to phase variants comprehensively within a single gene and its *cis*-regulatory regions. For example, in the context of autosomal recessive diseases with multiple segregating risk alleles, such as cystic fibrosis, ascertainment of local haplotypes covering the disease locus sheds light on compound heterozygosity and assists with clinical diagnosis. For these applications, targeted haplotype resolution technologies may be preferred owing to a reduction in both experimental costs (reagents, labour and turnaround time) and the burden of incidental findings (BOX 1). However, in other cases chromosome-scale haplotype resolution offers additional value. In some cancers, the phasing of distant somatic variants that have arisen on the same haplotype during tumour progression can be used in lineage tracing — for example,

Runs of homozygosity
Regions of the genome above a given distance threshold at which both haplotypes are identical.

Compound heterozygosity
The presence of two different recessive alleles, one on each haplotype, in a specific gene in a single individual. It is particularly relevant for autosomal recessive genetic diseases, which are frequently caused by compound heterozygosity in non-consanguineous pedigrees.

Variant imputation
A statistically grounded method for 'filling in' missing alleles in sparsely genotyped individuals to increase the power of association studies on the basis of similarity to reference panels of previously ascertained haplotypes.

by providing evidence that these linked variants are derived from the same clonal subpopulation. Whole-chromosome haplotype resolution is also valuable for the analysis of chromosome-wide regulatory mechanisms, such as long non-coding RNA-mediated control of X chromosome inactivation[79], or of autosomal replication timing[80].

Conclusions

Haplotype information is essential to the complete description of individual genomes. Many experimental methods for haplotype-resolved genome sequencing have recently been developed, but they differ in terms of comprehensiveness over variants and physical distances, accuracy and ease of implementation. As long-read sequencing technologies continue to mature, and as low-input amplification methods improve, we anticipate that these direct haplotyping methods will be refined over the next few years, becoming more cost-effective and increasingly adoptable to automation, multiplexing and routine use. As the number of dense haplotype-resolved genome sequences grows, hybrid methods designed to take advantage of this information are likely to gain stature. Particularly, as genomics achieves wider adoption in medicine, we predict that these haplotype-resolving technologies will be broadly adopted to maximize the completeness and utility of the human genomes that are sequenced.

1. Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genet.* **44**, 631–635 (2012).
2. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genet.* **45**, 1150–1159 (2013).
3. Tsoi, L. C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nature Genet.* **44**, 1341–1348 (2012).
4. Nalls, M. A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nature Genet.* **46**, 989–993 (2014).
5. Vernot, B. & Akey, J. M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**, 1017–1021 (2014).
6. Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354–357 (2014).
7. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature Genet.* **46**, 919–925 (2014).
8. Drysdale, C. M. *et al.* Complex promoter and coding region β_2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proc. Natl Acad. Sci. USA* **97**, 10483–10488 (2000).
9. Deenen, M. J. *et al.* Relationship between single nucleotide polymorphisms and haplotypes in *DPYD* and toxicity and efficacy of capecitabine in advanced colorectal cancer. *Clin. Cancer Res.* **17**, 3455–3468 (2011).
10. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nature Rev. Genet.* **12**, 703–714 (2011).
11. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **490**, 56–65 (2012).
12. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
13. Reich, D. *et al.* Reduced neutrophil count in people of African descent is due to a regulatory variant in the duffy antigen receptor for chemokines gene. *PLoS Genet.* **5**, e1000360 (2009).
14. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
15. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
16. Venter, J. C. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
17. Shendure, J. & Aiden, E. L. The expanding scope of DNA sequencing. *Nature Biotech.* **30**, 1084–1094 (2012).
18. McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009).
19. Dear, P. H. & Cook, P. R. Happy mapping: a proposal for linkage mapping the human genome. *Nucleic Acids Res.* **17**, 6795–6807 (1989).
This paper provides the conceptual framework for various subsequent phasing approaches that exploit the physical linkage between markers on HMW DNA and rely on limiting dilution to sub-haploid pools.
20. Burgtorf, C. *et al.* Clone-based systematic haplotyping (CSH): a procedure for physical haplotyping of whole genomes. *Genome Res.* **13**, 2717–2724 (2003).
This paper describes haplotype resolution using fosmid clone sequencing and laid the groundwork for massively parallel implementations.
21. Raymond, C. K. *et al.* Targeted, haplotype-resolved resequencing of long segments of the human genome. *Genomics* **86**, 759–766 (2005).
22. Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature Biotech.* **29**, 59–63 (2011).
This is the first report of a molecularly phased human genome that was sequenced on a massively parallel, short-read sequencing platform.
23. Lo, C. *et al.* On the design of clone-based haplotyping. *Genome Biol.* **14**, R100 (2013).
24. Suk, E. K. *et al.* A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res.* **21**, 1672–1685 (2011).
25. Adey, A. *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207–211 (2013).
26. Kitzman, J. O. *et al.* Noninvasive whole-genome sequencing of a human fetus. *Sci. Transl Med.* **4**, 137ra76 (2012).
27. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
28. Duitama, J. *et al.* Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of single individual haplotyping techniques. *Nucleic Acids Res.* **40**, 2041–2053 (2012).
29. Hoehe, M. R. *et al.* Multiple haplotype-resolved genomes reveal population patterns of gene and protein diplotypes. *Nature Commun.* **5**, 5569 (2014).
30. Paul, P. & Apgar, J. Single-molecule dilution and multiple displacement amplification for molecular haplotyping. *BioTechniques* **38**, 553–559 (2005).
31. Peters, B. A. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190–195 (2012).
This paper describes a fully *in vitro* approach for sequencing and phasing human genomes in a production setting with greatly reduced requirements for input DNA mass.
32. Kaper, F. *et al.* Whole-genome haplotyping by dilution, amplification, and sequencing. *Proc. Natl Acad. Sci. USA* **110**, 5552–5557 (2013).
33. Kuleshov, V. *et al.* Whole-genome haplotyping using long reads and statistical methods. *Nature Biotech.* **32**, 261–266 (2014).
34. Amini, S. *et al.* Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nature Genet.* **46**, 1343–1349 (2014).
35. Krol, A. 10X Genomics at AGBT. *Bio-ITWorld* [online], http://www.bio-itworld.com/2015/2/25/10x-genomics-agbt.html (2015).
36 Hiatt, J. B., Patwardhan, R. P., Turner, E. H., Lee, C. & Shendure, J. Parallel, tag-directed assembly of locally derived short sequence reads. *Nature Meth.* **7**, 119–122 (2010).
37. Voskoboynik, A. *et al.* The genome sequence of the colonial chordate, *Botryllus schlosseri*. *eLife* **2**, e00569 (2013).
38. Laszlo, A. H. *et al.* Decoding long nanopore sequencing reads of natural DNA. *Nature Biotech.* **32**, 829–833 (2014).
39. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
40. Yan, H. *et al.* Conversion of diploidy to haploidy. *Nature* **403**, 723–724 (2000).
41. Douglas, J. A., Boehnke, M., Gillanders, E., Trent, J. M. & Gruber, S. B. Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nature Genet.* **28**, 361–364 (2001).
42. Zhang, K. *et al.* Long-range polony haplotyping of individual human chromosome molecules. *Nature Genet.* **38**, 382–387 (2006).
43. Ma, L. *et al.* Direct determination of molecular haplotypes by chromosome microdissection. *Nature Meth.* **7**, 299–301 (2010).
44. Yang, H., Chen, X. & Wong, W. H. Completely phased genome sequencing through chromosome sorting. *Proc. Natl Acad. Sci. USA* **108**, 12–17 (2011).
45. Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nature Biotech.* **29**, 51–57 (2010).
46. Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626 (2012).
47. Wang, J., Fan, H. C., Behr, B. & Quake, S. R. Genome-wide single-cell analysis of recombination activity and *de novo* mutation rates in human sperm. *Cell* **150**, 402–412 (2012).
48. Kirkness, E. F. *et al.* Sequencing of isolated sperm cells for direct haplotyping of a human genome. *Genome Res.* **23**, 826–832 (2013).
49. Lu, S. *et al.* Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* **338**, 1627–1630 (2012).
50. Hou, Y. *et al.* Genome analyses of single human oocytes. *Cell* **155**, 1492–1506 (2013).
51. de Bourcy, C. F. A. *et al.* A quantitative comparison of single-cell whole genome amplification methods. *PLoS ONE* **9**, e105585 (2014).
52. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
53. Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465**, 363–367 (2010).
54. Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Rev. Genet.* **14**, 390–403 (2013).
55. Selvaraj, S., R. Dixon, J., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature Biotech.* **31**, 1111–1118 (2013).
This paper reports the first use of chromatin interaction maps to capture long-range sparse haplotypes along with a hybrid strategy to increase haplotype density.
56. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *arXiv* [online], http://arxiv.org/abs/1502.05331 (2015).

57. Lancia, G., Bafna, V., Istrail, S., Lippert, R. & Schwartz, R. in *Lecture Notes in Computer Science* Vol. 2161 (eds Goos, G. *et al.*)182–193 (Springer, 2001).
58. Bansal, V., Halpern, A. L., Axelrod, N. & Bafna, V. An MCMC algorithm for haplotype assembly from whole-genome sequence data. *Genome Res.* **18**, 1336–1346 (2008).
59. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**, i153–i159 (2008).
60. Aguiar, D. & Istrail, S. Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics* **29**, 352–360 (2013).
61. Aguiar, D. & Istrail, S. HapCompass: a fast cycle basis algorithm for accurate haplotype assembly of sequence data. *J. Comp. Bio.* **19**, 577–590 (2012).
62. Lo, C., Bashir, A., Bansal, V. & Bafna, V. Strobe sequence design for haplotype assembly. *BMC Bioinformatics* **12**, S24 (2011).
63. Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F. & Marchini, J. Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* **93**, 687–696 (2013).
64. Zhang, K. & Zhi, D. Joint haplotype phasing and genotype calling of multiple individuals using haplotype informative reads. *Bioinformatics* **29**, 2427–2434 (2013).
65. Yang, W. Y. *et al.* Leveraging reads that span multiple single nucleotide polymorphisms for haplotype inference from sequencing data. *Bioinformatics* **29**, 2245–2252 (2013).
66. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv* [online], http://arxiv.org/abs/1207.3907 (2012).
67. Fan, H. C. *et al.* Non-invasive prenatal measurement of the fetal genome. *Nature* **487**, 320–324 (2012).
68. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
69. Kuleshov, V. Probabilistic single-individual haplotyping. *Bioinformatics* **30**, i379–i385 (2014).
70. Matsumoto, H. & Kiryu, H. MixSIH: a mixture model for single individual haplotyping. *BMC Genomics* **14**, S5 (2013).
71. Pemberton, T. J. *et al.* Genomic patterns of homozygosity in worldwide human populations. *Am. J. Hum. Genet.* **91**, 275–292 (2012).
72. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
This paper demonstrates the use of allelic imbalance across long blocks of phased markers as a signal for aneuploidy in tumour genomes.
73. Schaaf, C. P. *et al.* Truncating mutations of *MAGEL2* cause Prader–Willi phenotypes and autism. *Nature Genet.* **45**, 1405–1408 (2013).
74. Wang, L. *et al.* Programming and inheritance of parental DNA methylomes in mammals. *Cell* **157**, 979–991 (2014).
75. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **488**, 91–100 (2012).
76. Xie, W. *et al.* Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148**, 816–831 (2012).
77. Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**, 350–354 (2015).
78. Lo, Y. M. D. *et al.* Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl Med.* **2**, 61ra91 (2010).
79. Brown, C. J. *et al.* The human *XIST* gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**, 527–542 (1992).
80. Stoffregen, E. P., Donley, N., Stauffer, D., Smith, L. & Thayer, M. J. An autosomal locus that controls chromosome-wide replication timing and mono-allelic expression. *Hum. Mol. Genet.* **20**, 2366–2378 (2011).
81. Xiao, M. *et al.* Determination of haplotypes from single DNA molecules: a method for single-molecule barcoding. *Hum. Mutat.* **28**, 913–921 (2007).
82. Xiao, M. *et al.* Direct determination of haplotypes from single DNA molecules. *Nature Meth.* **6**, 199–201 (2009).
83. Mitra, R. D. *et al.* Digital genotyping and haplotyping with polymerase colonies. *Proc. Natl Acad. Sci. USA* **100**, 5926–5931 (2003).
84. Wetmur, J. G. Molecular haplotyping by linking emulsion PCR: analysis of paraoxonase 1 haplotypes and phenotypes. *Nucleic Acids Res.* **33**, 2615–2619 (2005).
85. Turner, D. J. *et al.* Assaying chromosomal inversions by single-molecule haplotyping. *Nature Meth.* **3**, 439–445 (2006).
86. Regan, J. F. *et al.* A rapid molecular approach for chromosomal phasing. *PLoS ONE* **10**, e0118270 (2015).
87. Nedelkova, M. *et al.* Targeted isolation of cloned genomic regions by recombineering for haplotype phasing and isogenic targeting. *Nucleic Acids Res.* **39**, e137 (2011).
88. de Vree, P. J. P. *et al.* Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nature Biotech.* **32**, 1019–1025 (2014).
89. Adey, A. *et al. In vitro*, long-range sequence information for *de novo* genome assembly via transposase contiguity. *Genome Res.* **24**, 2041–2049 (2014).
90. Steinberg, K. M. *et al.* Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res.* **24**, 2066–2076 (2014).
91. Bauer, D. E. *et al.* An erythroid enhancer of *BCL11A* subject to genetic variation determines fetal hemoglobin level. *Science* **342**, 253–257 (2013).

Acknowledgements

The authors thank B. Browning, B. Vernot, A. Gordon and members of the Shendure Lab for discussions.

Competing interests statement

The authors declare competing interests: see Web version for details.