Journal of Assisted Reproduction and Genetics (2019) 36:727–739
https://doi.org/10.1007/s10815-018-1392-1

GENETICS

Off the street phasing (OTSP): no hassle haplotype phasing for molecular PGD applications

David A. Zeevi[1] · Fouad Zahdeh[1] · Yehuda Kling[1] · Shai Carmi[2] · Gheona Altarescu[1]

Received: 15 October 2018 / Accepted: 18 December 2018 / Published online: 8 January 2019

Abstract

Purpose Pre-implantation genetic diagnosis (PGD) for molecular disorders requires the construction of parental haplotypes. Classically, haplotype resolution ("phasing") is obtained by genotyping multiple polymorphic markers in both parents and at least one additional relative. However, this process is time-consuming, and immediate family members are not always available. The recent availability of massive genomic data for many populations promises to eliminate the needs for developing family-specific assays and for recruiting additional family members. In this study, we aimed to validate population-assisted haplotype phasing for PGD.

Methods Targeted sequencing of *CFTR* gene variants and ~ 1700 flanking polymorphic SNPs (± 2 Mb) was performed on 54 individuals from 12 PGD families of (a) Full Ashkenazi (FA; $n = 16$), (b) mixed Ashkenazi (MA; $n = 23$ individuals with at least one Ashkenazi and one non-Ashkenazi grandparents), or (c) non-Ashkenazi (NA; $n = 15$) descent. Heterozygous genotype calls in each individual were phased using various whole genome reference panels and appropriate computational models. All computationally derived haplotype predictions were benchmarked against trio-based phasing.

Results Using the Ashkenazi reference panel, phasing of FA was highly accurate (99.4% ± 0.2% accuracy); phasing of MA was less accurate (95.4% ± 4.5% accuracy); and phasing of NA was predictably low (83.4% ± 6.6% accuracy). Strikingly, for founder mutation carriers, our haplotyping approach facilitated near perfect phasing accuracy (99.9% ± 0.1% and 98.2% ± 2.8% accuracy for W1282X and delF508 carriers, respectively).

Conclusions Our results demonstrate the feasibility of replacing classical haplotype phasing with population-based phasing with uncompromised accuracy.

Keywords PGD · Haplotype phasing · Population-based phasing · Identity by descent · CFTR

Introduction

Pre-implantation genetic diagnosis (PGD) is performed for couples at high genetic risk for Mendelian and chromosomal disorders by a biopsy of the blastomere, blastocyst, and/or the polar body [1]. The main problem hampering the accuracy of diagnosis in single-cell analysis remains the occurrence of allele dropout (ADO), namely random amplification failure of one or both alleles. This can lead to misdiagnosis, by mistakenly calling heterozygous genotypes as homozygous, or to no diagnosis. Therefore, testing of several polymorphic markers flanking the familial mutation is essential for identification of the transmitted parental haplotypes and, thereby, for increasing accuracy. Complicating the procedure is that reconstruction of the transmitted haplotypes requires, as a prerequisite, phasing the haplotypes of the parents.

Classically, haplotype phasing is determined empirically by genotyping both the parents and at least one of their first degree relatives (with preference for a child or for both sets of parents of each partner). However, immediate family members are not always available, and in some cases, the couple does not agree to involve family members. Another obstacle is polymorphic marker selection, which is family-specific (due to the need to identify informative polymorphic markers in each couple) and thus generally time-consuming and

✉ David A. Zeevi
zeevidavid@szmc.org.il

[1] Medical Genetics Institute, Shaare Zedek Medical Center (SZMC), Bayit Str. 12, P.O.Box 3235, 91031 Jerusalem, Israel

[2] Braun School of Public Health and Community Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel

laborious. As a result, there is a relatively long delay between the couple's first meeting for genetic counseling and the initiation of in vitro fertilization (IVF) and PGD. To counter family-specific marker selection, whole genome haplotyping methods such as karyomapping and haplarithmesis have been developed recently for PGD application [2–12]. However, these methods still require immediate family members of the couple to facilitate accurate whole genome phasing of the embryo and so they too are limited by the availability of DNA from individuals who are not always accessible to the PGD clinic. Therefore, there is a clear need to develop new phasing methods that do not require family member recruitment beyond that of the couple seeking PGD.

Recently, large-scale data sets of whole-genome sequences have become available for thousands of individuals across numerous populations [13–15]. These "reference" genomes, along with sophisticated statistical methods, provide means for highly accurate haplotype phasing even in the absence of close relatives, by taking advantage of remote relatedness between the target genomes and reference haplotypes [16–18]. Several studies have shown that the accuracy of "population-based" phasing increases rapidly with the size of the reference data set and that accuracy further increases in correlation with increasing ethnic relatedness between the reference and the target genomes [17–22].

A number of population-based phasing approaches work by modeling haplotype frequencies and using them to estimate the most probable haplotype configuration for a target genome (e.g., BEAGLE [22, 23]). Most recent methods are based on an approximate coalescent model, whereby the haplotype to be inferred is assumed to be an imperfect mosaic of short-range "reference" haplotype blocks from the population. Transitions between reference haplotypes follow historical recombinations and are inferred using a hidden Markov model (e.g., MACH [24], IMPUTE2 [25], SHAPEIT [26–28], HAPI-UR [20], EAGLE2 [18]).

A number of other methods are based on the idea of long-range phasing [21], which is conceptually similar to phasing related individuals. For close relatives, a number of megabase-long haplotypes are identical-by-descent (IBD) between the two individuals. In these regions, phasing is trivial at sites where at least one individual is homozygous. For unrelated individuals from the same population, long IBD segments (e.g., > 5 cM) are also present in large numbers, in particular in founder populations, representing haplotypes transmitted from common ancestors who have lived relatively recently (a few tens of generations ago). Long-range phasing is usually rule-based rather than model-based and proceeds by first identifying IBD segments (based on high allelic similarity across a long region) and then using various techniques to infer the haplotypes in these regions and propagate the information across individuals [29, 30] [16, 21, 31]. Here, we used SHAPEIT2 as the population-based short-range phasing tool, and we built our own IBD-based phasing method using a nearest neighbor heuristic.

Our goal was to assess the feasibility of replacing traditional family-based haplotype phasing with less resource-demanding population-based phasing (whether short- or long-ranged). Specifically, we attempted to evaluate population-assisted haplotype phasing in a hospital-based pre-clinical setting in Israel, where the patient population is composed of both Ashkenazi Jewish (AJ) individuals, for which extensive whole-genome sequencing data is available, as well as individuals of other ancestries. We assessed the effects of ethnicity-matching, reference panel population size, and short- vs long-range phasing on haplotype prediction accuracy. The results show that given proper circumstances and common pre-conditions, population-based phasing can indeed eliminate the need to recruit relatives and/or design family-specific assays.

Materials and methods

Sample collection

DNA samples were collected from 54 individuals from 12 PGD families as part of routine pre-case workup at the Shaare Zedek Medical Center PGD lab. The study population consisted of three ethnic subgroups: (a) full Ashkenazi Jewish ("FA"; individuals whose both parents have an AJ genetic origin; $n = 16$); (b) mixed Ashkenazi ("MA"; individuals with at least one Ashkenazi grandparent and at least one non-Ashkenazi grandparent; $n = 23$); and (c) non-Ashkenazi ("NA"; individuals without any Ashkenazi grandparent; $n = 15$). Eight W1282X and four delF508 *CFTR* founder mutation carriers were among the FA and MA subgroups.

High throughput targeted sequencing

A custom next-generation sequencing panel was designed to target 1740 common *CFTR* variants and gene-flanking polymorphic SNPs (± 2 Mb from *CFTR*) with minor allele frequency > 25%. Briefly, a pool of 1740 primer pairs was used to PCR amplify the targeted *CFTR*-flanking SNPs in each DNA sample using the GeneRead DNAseq Panel PCR Kit V2 (Qiagen) according to the manufacturer's protocol. Subsequently, multiplex PCR products from each sample were converted into indexed high throughput sequencing libraries using the QIAseq 1-Step Amplicon Library Kit (Qiagen), also according to the manufacturer's protocol. Following library prep, indexed samples were pooled and sequenced on MiSeq or NextSeq 500 instruments (Illumina) to 1000× mean coverage per PCR amplicon.

Population-based short-range haplotype phasing

We phased all samples using SHAPEIT v2.r837 [32], using unrelated reference genomes [27]. Reference panels were either the 1000 Genomes Project (1092 samples) [33] or whole genome sequences of either 128 [34] or 574 [35] Ashkenazi Jewish individuals (with no overlap between the two panels), both provided by The Ashkenazi Genome Consortium. The haplotypes were compared against ground-truth phasing, which we obtained, for each sample, using trio information.

Haplotype clustering

To visualize *CFTR* delF508 and W1282X mutation carrier haplotypes, we used a UPGMA (Unweighted Pair Group Method with Arithmetic mean) agglomerative hierarchal clustering model. Haplotypes were first represented as binary sequences (0: reference allele, 1: non-reference allele). Hamming distances were then computed between every pair of haplotypes. Clusters of single haplotypes that were in close proximity, based on their average pairwise distance, were repeatedly grouped/linked into larger clusters until a hierarchical tree was formed.

IBD-based phasing

Our nearest neighbor heuristic-based IBD-phasing method starts from a set of reference (phased) haplotypes. To phase heterozygote sites in a target genome, we first masked these sites and used the remaining homozygote sites to determine which reference haplotype is the nearest. To quantify similarity between haplotypes, we used the Hamming distance, when encoding the reference alleles as 0 and the alternative alleles as 1. For each mutation carrier, we identified the nearest reference haplotype and phased the (previously masked) heterozygote sites according to that reference (namely, by assigning the allele that appears in the reference to the mutation-carrying haplotype).

Phasing of CFTR founder mutation carriers

To improve phasing of haplotypes carrying *CFTR* founder mutations, we developed mutation-specific sets of reference haplotypes, as follows. For the *CFTR* delF508 founder mutation, we first performed targeted amplicon sequencing of the same 1740 *CFTR*-flanking polymorphic SNPs (as described above) on a new in-house sample of 15 carrier trios spanning multiple ethnicities. We then merged the genomes from the 15 trios with those of the 574AJ reference panel (along SNPs that appear in both data sets) and phased delF508 carrier FA samples using SHAPEIT with the new expanded reference panel. In addition, we also determined ground truth delF508-linked haplotypes from each lab reference trio based on familial information. For each such trio, we kept only the haplotype carrying the founder mutation to generate a reference set of 15 carrier haplotypes for IBD phasing. Then, to phase heterozygote sites in the delF508 carriers, we used the above mentioned nearest neighbor IBD-phasing method with either the 15 carrier trios or the 574AJ delF508 carriers (4 samples) as the reference.

For phasing W1282X mutation carriers, we used the same nearest neighbor heuristic using W1282X carriers (and noncarriers for control) from the 574 reference AJ data set. Construction of a W1282X-specific lab reference was unnecessary to improve haplotype phasing accuracy (see "Results").

Benchmarking of haplotype phasing methods

To benchmark the performance of population-based phasing methods, we compared the resulting haplotypes from all methods to those computed based on Mendelian inheritance in the trio sequencing data. Phasing accuracy was defined as the proportion of SNPs where the inferred maternal and paternal haplotypes were the same as based on Mendelian inheritance. Phasing accuracy was compared across the different analyses using paired *t* test. *P* values were Bonferroni-corrected for multiple comparisons.

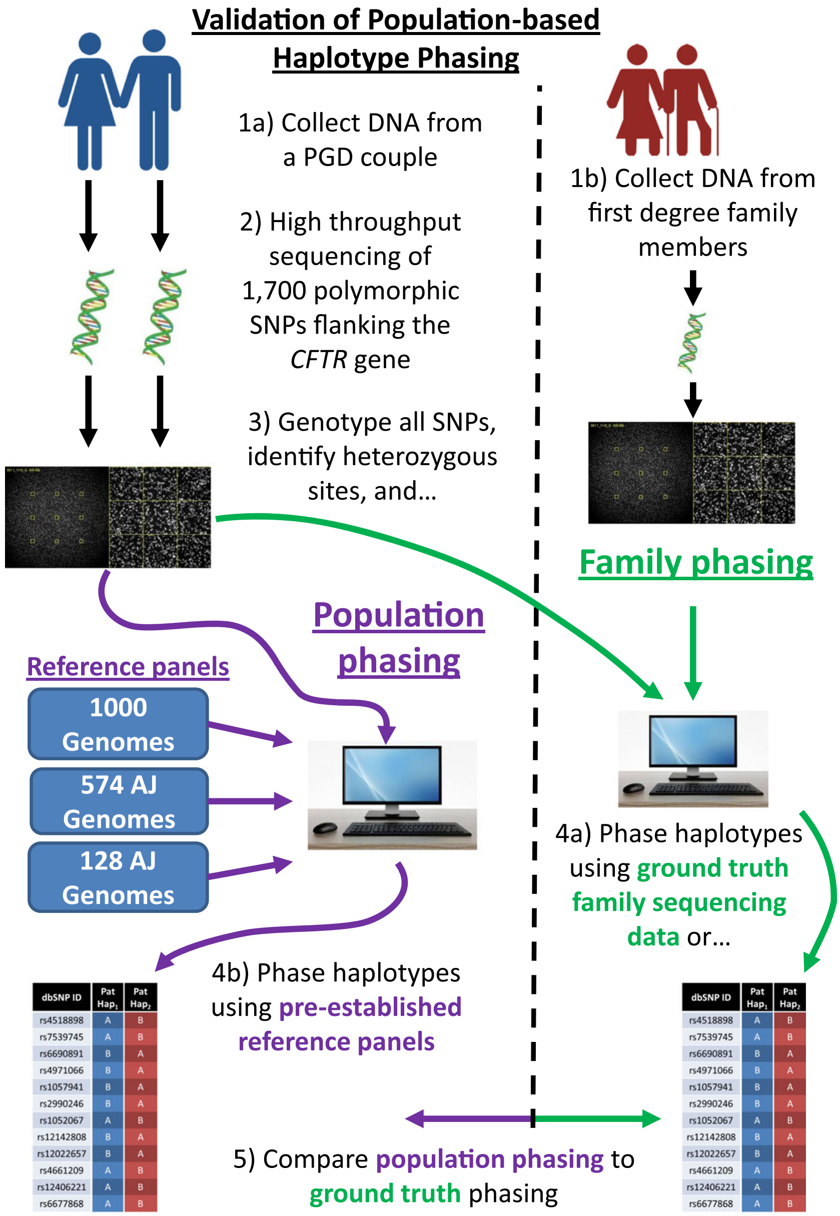
Results

Accurate population-based haplotype phasing of full and mixed Ashkenazi individuals

We sampled 54 individuals from 12 PGD families and used three population reference panels (1000 Genomes, 128 AJ genomes, and 574 AJ genomes) for population-based haplotype phasing of the *CFTR* genomic region in each of the individuals (Fig. 1). Given that many of our samples were first degree relatives, we could validate population-based phasing (of unrelated individuals) against haplotypes derived directly based on familial relationships.

The primary purpose of analyzing the three reference panels was to assess the effect of ethnicity-matching and panel size on phasing accuracy. For ethnicity matching, we compared the diverse 1000 Genomes reference panel (1000G; comprised of many worldwide ethnicities) to more study-appropriate ethnic matched Ashkenazi Jewish reference panels. For reference panel size assessment we compared the relatively large 1000G and 574 AJ genome reference panels with the relatively small 128 AJ genome reference panel.

The ethnic background and *CFTR* mutation carriage (where relevant) among the study cohort are detailed in Table 1. For population-based phasing, we treated all individuals as "unrelated" in the bioinformatic phasing pipeline. The samples were divided by ethnicity into three subgroups: (1) full


Validation of Population-based Haplotype Phasing
1a) Collect DNA from a PGD couple
2) High throughput sequencing of 1,700 polymorphic SNPs flanking the *CFTR* gene
3) Genotype all SNPs, identify heterozygous sites, and…
1b) Collect DNA from first degree family members
Family phasing
Population phasing
Reference panels
1000 Genomes
574 AJ Genomes
128 AJ Genomes
4a) Phase haplotypes using ground truth family sequencing data or…
4b) Phase haplotypes using pre-established reference panels
5) Compare population phasing to ground truth phasing

dbSNP ID	Pat Hap_1	Pat Hap_2
rs4518898	A	B
rs7539745	A	B
rs6690891	B	A
rs4971066	B	A
rs1057941	B	A
rs2990246	B	A
rs1052067	A	B
rs12142808	B	A
rs12022657	B	A
rs4661209	A	B
rs12406221	A	B
rs6677868	A	B

dbSNP ID	Pat Hap_1	Pat Hap_2
rs4518898	A	B
rs7539745	A	B
rs6690891	B	A
rs4971066	B	A
rs1057941	B	A
rs2990246	B	A
rs1052067	A	B
rs12142808	B	A
rs12022657	B	A
rs4661209	A	B
rs12406221	A	B
rs6677868	A	B

Fig. 1 Our strategy for validation of population-based haplotype phasing. Statements (1) through (5) describe sequential steps of the haplotype phasing process. Items to the left of the segmented vertical line (highlighted with purple text and arrows) illustrate population phasing-specific processes. Items to the right of the segmented vertical line (highlighted with green text and arrows) illustrate ground truth family phasing-specific processes. The aim of these experiments was to assess whether haplotypes of a PGD couple could be accurately reconstructed without requiring the DNA sequence of their first degree relatives and/or tailored genetic assay

Ashkenazi (FA; $n = 16$), (2) mixed Ashkenazi (MA; $n = 23$), and (3) non-Ashkenazi (NA; $n = 15$). High throughput sequencing of ~ 1700 *CFTR*-flanking SNPs (± 2 Mb from *CFTR*) was performed on all samples, and heterozygous SNPs were identified for haplotype phasing. Overall, we found an average of ~ 540 heterozygous SNPs per sample across the *CFTR* gene-flanking region, with similar heterozygosity across all subgroups and similar SNP marker coverage across the three reference panels (1000G, 128 AJ, and 574 AJ).

We show the phasing accuracy across the various experimental conditions in Fig. 2. In the FA subgroup (Fig. 2a), although the 1000G data set was the largest panel, it resulted in the lowest phasing accuracy (66% ± 4.7%). Significantly better results were observed when using the small ($n = 128$) ethnicity-matched AJ reference panel (90.4% ± 5.7% phasing accuracy; $P = 3.0 \times 10^{-2}$ in comparison to 1000G; Bonferroni-corrected paired Student's *t* test). Haplotype accuracy was the best when using a larger AJ panel of $n = 574$ (99.4% ± 0.2%; $P = 0.42$ for 574 AJ vs 128 AJ and $P = 3.0 \times 10^{-4}$ vs 1000G; Bonferroni-corrected paired *t* test). Therefore, these results suggest that ethnicity-matching, combined with large reference panel size, can lead to nearly perfect population-based haplotype phasing.

To provide context to the above results, we performed population-based phasing on 15 "negative control" NA samples. As these individuals lack AJ ancestry completely, we expected that they would not be "phasable" by AJ reference

Table 1 Study recruit ethnicity and *CFTR* mutation carriage, grouped according to family number

Sample[a]	Family no.	Ethnicity	Ethnicity classification	*CFTR* genotype[b]
Ch19032	0	AJ/Kurdish[c]	MA	W1282X/3121-1G>A
F19544	0	Kurdish/Algerian	NA	3121-1G>A/wt
M19543	0	AJ/Turkish/Iraqi	MA	W1282X/wt
Ch36691	1	AJ/Iraqi/Turkish	MA	wt/wt
F26423	1	Iraqi/Turkish	NA	wt/wt
GFF26560	1	Iraqi/Turkish	NA	wt/wt
GMF28650	1	Iraqi/Turkish	NA	wt/wt
M26422	1	AJ	FA	wt/wt
Ch31028	2	Palestinian Arab	NA	wt/wt
Ch38583	2	Palestinian Arab	NA	wt/wt
F30494	2	Palestinian Arab	NA	wt/wt
M30493	2	Palestinian Arab	NA	wt/wt
Ch10793	3	AJ/Persian	MA	wt/wt
Ch10794	3	AJ/Persian	MA	wt/wt
Ch38901	3	AJ/Persian	MA	wt/wt
F8680	3	AJ/Persian	MA	wt/wt
M8679	3	AJ	FA	wt/wt
Ch22695	4	AJ/Tunisian	MA	E819X/N1303K
F23116	4	Tunisian	NA	E819X/wt
M23920	4	AJ	FA	N1303K/wt
Ch25480	5	AJ/Tripoli	MA	wt/wt
Ch36735	5	AJ/Tripoli	MA	wt/wt
F20376	5	AJ/Tripoli	MA	W1282X/wt
M20375	5	AJ	FA	delF508/wt
F33713	6	AJ	FA	W1282X/wt
GFF33717	6	AJ	FA	W1282X/wt
GFM33714	6	AJ	FA	delF508/wt
GMF33716	6	AJ	FA	wt/wt
GMM33715	6	AJ	FA	wt/wt
M33712	6	AJ	FA	delF508/wt
F33859	7	AJ	FA	delF508/wt
GFM33910	7	Tunisian	NA	405 + 1/wt
GMF33730	7	AJ	FA	wt/wt
GMM33911	7	AJ	FA	wt/wt
M33848	7	AJ/Tunisian	MA	405 + 1/wt
Ch30004	8	AJ/non-Jewish	MA	wt/wt
Ch30005	8	AJ/non-Jewish	MA	wt/wt
Ch38446	8	AJ/non-Jewish	MA	wt/wt
F29887	8	AJ	FA	wt/wt
M29886	8	AJ/non-Jewish	MA	wt/wt
F8420	9	Turkish	NA	wt/wt
GFF8457	9	Turkish	NA	wt/wt
GFM8455	9	Iraqi	NA	wt/wt
GMF8456	9	Turkish	NA	wt/wt
M8419	9	AJ/Iraqi	MA	wt/wt
Ch19153	11	AJ/Moroccan	MA	W1282X/wt
Ch19154	11	AJ/Moroccan	MA	W1282X/3849 + 10 kb
F19000	11	AJ	FA	W1282X/wt
M18999	11	Moroccan	NA	3849 + 10 kb/wt

Table 1 (continued)

Sample[a]	Family no.	Ethnicity	Ethnicity classification	*CFTR* genotype[b]
Ch25724	13	AJ/Moroccan	MA	wt/wt
Ch25725	13	AJ/Moroccan	MA	wt/wt
Ch40315	13	AJ/Moroccan	MA	wt/wt
F25723	13	AJ	FA	wt/wt
M25722	13	AJ/Moroccan	MA	wt/wt

AJ Ashkenazi Jewish, *NA* non-Ashkenazi, *MA* mixed Ashkenazi Jewish, *FA* full Ashkenazi Jewish, *wt* wild type allele

[a] Family relationships are indicated by the leading alphabetic characters in the "Sample name." "Ch" indicates child; "F" indicates father; "M" indicates mother; "GFF" indicates paternal grandfather of the child (father side of the family); "GMF" indicates paternal grandmother of the child (father side of the family); "GFM" indicates maternal grandfather of the child (mother side of the family); "GMM" indicates maternal grandmother of the child (mother side of the family)

[b] *CFTR* mutations are provided according to legacy designation

[c] Ethnicity for this sample is shown according to the known ethnic origins of the mutant alleles in the indicated *CFTR* genotype

panels. Indeed, the phasing accuracy in this subgroup was mediocre with both the 128 and 574 AJ panels (68.1% ± 5.4% and 83.4% ± 6.6%, respectively) and neither is significantly different from the phasing accuracy of the 1000G reference (68.5% ± 4.6%) (Fig. 2b). In contrast, the mixed AJ subgroup exhibited a similar trend to that of the FA subgroup (phasing accuracy with 1000G 74.2% ± 6.6, 128AJ 87.2% ± 4.2, 574AJ 95.4% ± 4.6) albeit without statistically significant differences between the three reference panels (Fig. 2c). These results suggest that as long as at least one haplotype of the target individual appears in the reference (or is similar to a haplotype from the reference), phasing accuracy can be high.

Haplotype phasing of CFTR founder mutation carriers

Within the study cohort were 12 individuals carrying a *CFTR* founder mutation. Eight carried the W1282X variant, which is the most common pathogenic allele in the AJ population, and four carried the delF508 variant, which is common in various pan-ethnic populations (most notably people of European descent). Importantly, both pathogenic variants were present (in small number) within the 128 AJ and 574 AJ reference panels. The delF508 variant is also represented in the 1000G data set. Among the W1282X carriers, five were of MA ethnicity and three were FA. All four delF508 carriers were of FA ethnicity. Therefore, we expected that population phasing of all mutation carriers with the large 574 AJ reference panel would provide superior phasing accuracy compared to other FA. However, this was not always true.

Despite the fact that all four delF508 carriers were FA, only two individuals were phased by the 574AJ reference with high accuracy (> 99%, Table 2), while the other two were phased with accuracy of only 90.2% and 68.9%. Among the W1282X carriers, 574 AJ phasing accuracy was > 98.8% in all three FA

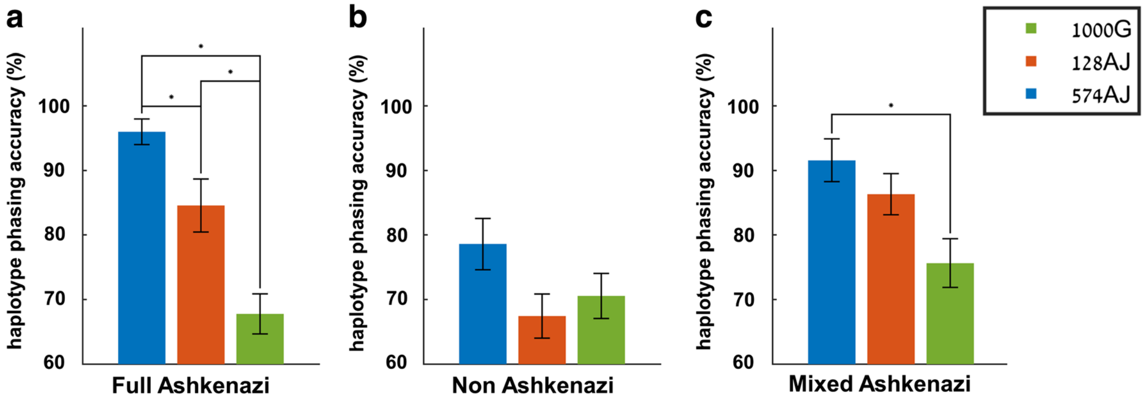


Fig. 2 Population haplotype phasing accuracy as a function of reference panel content and size. High throughput sequencing of ~1700 *CFTR*-flanking SNPs (± 2 Mb from *CFTR*) was performed on 54 samples from 12 families. We phased these individuals using a population-based approach (treating them as "unrelated"), with the aid of one of three reference panels. The reference panels were as follows: 574 unrelated AJ individuals (574 AJ), 128 unrelated AJ individuals (128 AJ) [34], and the publicly available 1000G [33]. Hundreds of heterozygous SNPs were identified in all samples. The figure shows the mean phasing accuracy (± SEM) for **a** full AJ ($n = 16$ individuals), **b** non-AJ ($n = 15$ individuals), and **c** mixed AJ ($n = 23$ individuals). The accuracy was determined by comparing the haplotypes against ground truth trio-phased haplotypes (Fig. 1). Asterisks indicate significant difference ($P < 0.05$; Bonferroni-corrected paired t test)

samples (Table 2). For MA samples, in three W1282X carriers, the phasing accuracy was > 99.9%, but for two others, accuracy was only 81.6% and 64.5% (Table 2).

Such unpredictable phasing accuracy is not suitable for clinical PGD application. Therefore, we hypothesized that the fluctuating accuracy was due to the relatively small number of carriers in the reference panels, which did not capture the entire diversity of haplotype sequences of carriers. However, for W1282X, this was clearly not the case, because there were 13 W1282X carrier alleles in the 574AJ reference, who all shared a single unified mutation-flanking haplotype (within ± 1 Mb of *CFTR*) not seen in non-carrier samples from the same reference data set (Fig. 3a).

On the other hand, for delF508, there was strong evidence for a limited coverage of haplotype diversity within the 574AJ reference. Indeed, there were only four delF508 carriers in the reference, and hierarchical clustering of their predicted delF508-linked haplotypes relative to non-delF08 alleles indicated the presence of two haplotypes among the four carriers (Fig. 3b). To further confirm this diversity of delF08 alleles in the AJ population, we performed targeted sequencing of *CFTR*-flanking SNPs (the same SNPs as those analyzed in the 54 sample study cohort) on an additional collection of in-house DNA samples from delF508 carriers. This lab-specific reference consisted of 15 delF508 carrier trios of AJ (9 trios) and non-AJ (6 trios for control) ancestry among whom none had any blood relationship to members of the study cohort. Here too, hierarchical clustering of the lab reference delF508 alleles identified more than one delF508-linked haplotype among AJ carrier alleles (Fig. 3c). Therefore, we surmised that correct haplotype phasing of W1282X and delF508 carriers, respectively, would require two different strategies to account for the lack of (in the case of W1282X) or presence of (in the case of delF508) mutation-linked allelic diversity in the AJ population.

To improve W1282X phasing, a new phasing strategy was devised to correct the haplotype predictions of two W1282X carriers of MA descent (samples M19543 and Ch19153) who were poorly phased by haplotype frequency-based phasing using the 574AJ reference (Table 2). In these cases, we expected that the presence of a non-AJ haplotype in each of the MA carriers would reduce phasing accuracy of short-range phasing algorithms such as SHAPEIT. Therefore, given the aforementioned homogeneity of W1282X alleles in the AJ population, we re-phased all W1282X mutation carriers in the study cohort using a strategy based on identity-by-descent (IBD) similarity between the target and reference genomes [21]. Specifically, using a simple nearest neighbor heuristic (see "Materials and methods"), the W1282X haplotypes in the 574AJ reference that best matched the profile of all

Table 2 The effect of population-based and nearest neighbor phasing on haplotype phasing of *CFTR* W1282X and delF508 mutation carriers in the study cohort

Sample	Ethnicity	Ethnicity classification	CFTR mutation	Phasing accuracy (%)					
				574AJ	128AJ	1000G	574AJ-IBD	LSR-IBD	574AJ+ LSR
M20375	Ashkenazi	FA	delF508/wt	99.7	99.4	83.8	82.7	99.0	99.5
GFM33714	Ashkenazi	FA	delF508/wt	99.1	79.6	68.2	53.2	94.9	99.3
M33712	Ashkenazi	FA	delF508/wt	90.2	67.0	72.7	77.6	97.2	99.8
F33859	Ashkenazi	FA	delF508/wt	68.9	68.9	56.1	66.9	96.4	94.0
FA delF508 average accuracy				*89.5*	*78.7*	*70.1*	*70.1*	*96.9*	*98.2*
Sample	Ethnicity	Ethnicity classification	CFTR mutation	Phasing accuracy (%)					
				574AJ		128AJ	1000G		574AJ-IBD
GFF33717	Ashkenazi	FA	W1282X/wt	99.5		61.0	57.7		100
F33713	Ashkenazi	FA	W1282X/wt	99.6		96.4	76.7		100
F19000	Ashkenazi	FA	W1282X/wt	98.8		98.8	52.3		99.5
FA W1282X average accuracy				*99.3*		*85.4*	*62.3*		*99.8*
Ch19154	Ashkenazi + Moroccan	MA	W1282X/3849 + 10 kb	100		86.9	75.5		100
F20376	Ashkenazi/Tripoli	MA	W1282X/wt	100		95.2	95.4		99.8
Ch19032	Kurdish + Ashkenazi	MA	3121/W1282X	99.9		92.0	64.5		99.8
M19543	Ashkenazi/Turkish + Iraqi	MA	W1282X/wt	81.6		100	58.3		100
Ch19153	Ashkenazi + Moroccan	MA	W1282X/wt	64.5		97.1	68.2		100
MA W1282X average accuracy				*89.2*		*94.2*	*72.4*		*99.9*

FA full Ashkenazi Jewish, *MA* mixed Ashkenazi Jewish, *wt* wild-type allele, *574AJ-IBD* identity-by-descent phasing using mutation carriers from the 574AJ reference data set, *LSR-IBD* identity-by-descent phasing using mutation carriers from the lab-specific reference data set, *574AJ+ LSR* population-phasing using the delF508 lab-specific trios combined with the 574AJ reference

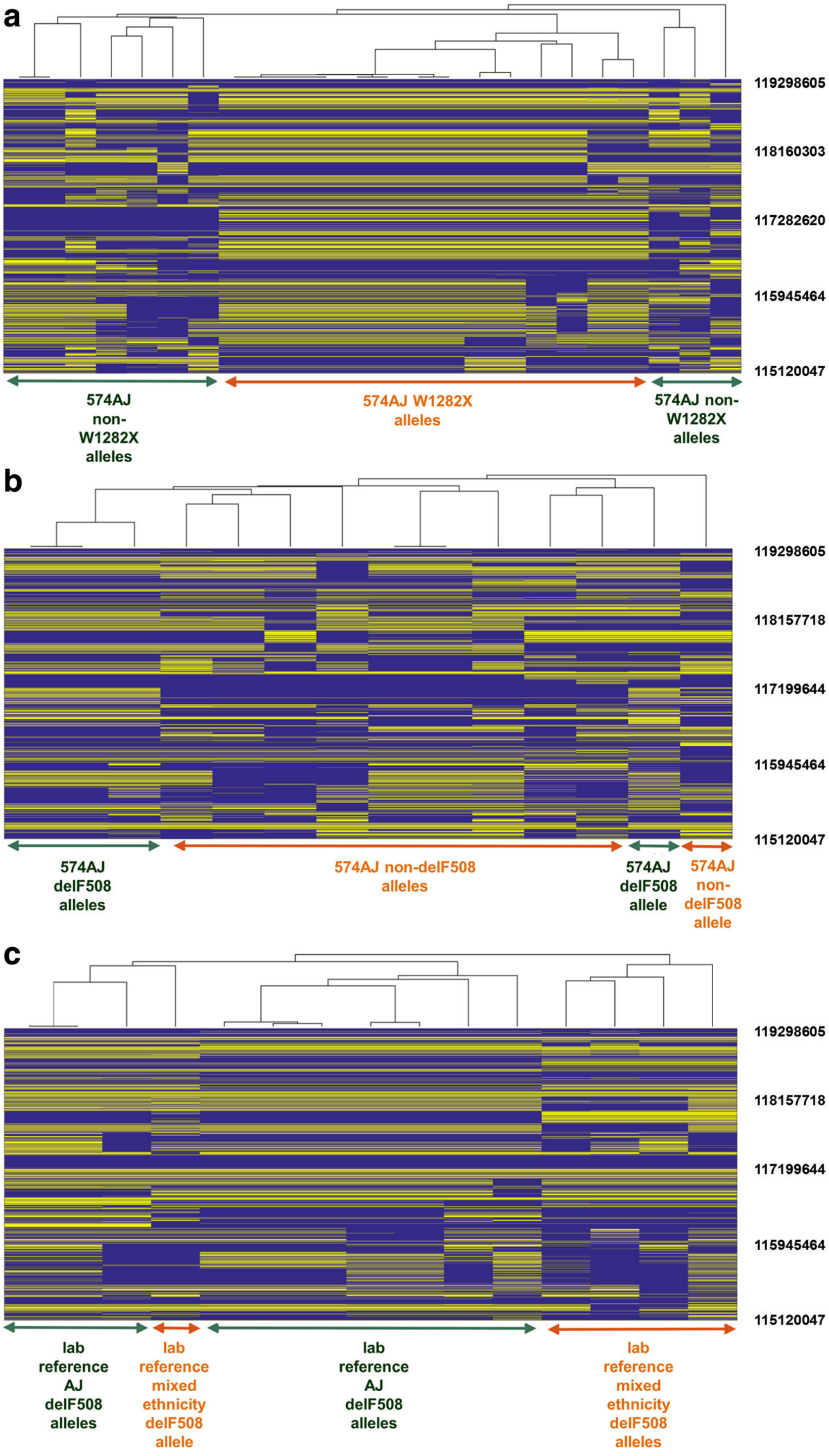


Fig. 3 Hierarchical clustering of W1282X and delF508 mutation-linked haplotypes identifies a single founder haplotype for the W1282X variant and multiple founder haplotypes for the delF508 variant. Pictured are dendrograms, generated by hierarchical clustering, of ~1700 sequenced SNPs flanking the *CFTR* gene that were phased into haplotype blocks. Purple coloring indicates an hg19 reference nucleotide and yellow coloring indicates the alternative allele. The physical position on hg19 chromosome 7 is indicated to the right. The *CFTR* gene is positioned in the horizontal center of each plot. **a** A dendrogram of 13 haplotypes who carry the *CFTR* W1282X allele in the 574 AJ reference data, together with 10 haplotypes carrying random non-W1282X alleles from the same reference set. Note the presence of a distinct W1282X-specific linkage disequilibrium block (not present in non-carrier controls) within ± 1 Mb of the *CFTR* gene in the middle of the 574AJ W1282X plot. **b** A dendrogram of four CFTR delF508-linked haplotypes in the 574 AJ reference data, together with 10 random non-delF508 CFTR alleles from the same reference data set. Note that not all four delF508 alleles cluster together. **c** A dendrogram of 15 CFTR delF508-linked haplotypes from our lab-specific reference, comprising nine AJ and six non-AJ mixed ethnicity (a Turkish-Romanian-AJ, a Turkish-Syrian-Caucasian, a Turkish-Persian, a Tunisian, and two Palestinian Arabs). The lab-specific reference was derived from amplicon targeted sequencing of ~1700 *CFTR*-flanking SNPs. Note that not all AJ delF508 alleles cluster together

homozygous sites of the mutation carrier were used to phase heterozygous SNP sites in that carrier. Strikingly, this strategy improved phasing accuracy of MA mutation carriers M19543 and Ch19153 (as well as other W1282X mutation carriers in the study cohort) to near perfection (Table 2; Fig. 4a). These results indicate that where a comprehensive mutation-specific reference is available (such as that for W1282X in the 574AJ reference data), IBD-based phasing should be considered the computational phasing method of choice to define gene-flanking haplotypes in mutation carriers of mixed ethnicity.

As described above, the delF508 mutation exists with more allelic diversity in the AJ population than W1282X. For this reason, we hypothesized that even delF508 carriers of FA descent could not be correctly phased using standard population-phasing strategies (Table 2). Indeed, even IBD-phasing, using the four delF508 carrier alleles in the 574AJ reference could not correctly identify the true haplotypes in all of the study cohort delF508 carriers (Table 2). This was evident from the fact that hierarchal clustering (see "Materials and methods") mis-categorized some FA delF508 carriers in the study cohort (such as sample F33859) together with non-delF508 haplotypes in the 574AJ reference, rather than true carrier haplotypes in the same reference data set (Fig. 4b). Thus, we sought to capture a more accurate representation of delF508-linked alleles in the AJ population by deriving our own delF508-specific lab reference for population-based and/or IBD-phasing (see "Materials and methods"). This lab reference was more informative for IBD than the 574AJ data because it contained more FA delF508 haplotypes, and it was also derived from ground truth trio-phasing. Accordingly, IBD-phasing with our delF508 lab reference provided much more reliable haplotype predictions for delF08 carriers in the study cohort than the other phasing strategies attempted above (96.9% mean lab-specific reference IBD-phasing accuracy vs 70.1% and 89.5% mean 574AJ IBD-phasing and mean 574AJ population-phasing accuracy, respectively; Table 2; Fig. 4c). This suggests that the lab reference contained a more inclusive representation of delF508 alleles than the original 574AJ reference.

We next hypothesized that by combining the lab-reference delF508 carriers with the 574AJ data, we could better capture the maximal number of relevant AJ alleles necessary to phase unknown FA delF508 carriers. Indeed, when we re-ran SHAPEIT with the 574AJ and lab delF508 samples combined into one all-inclusive reference, we observed the highest phasing accuracy of all (Table 2). Hence, to meet standard expectations of a clinical application for diverse haplotypes that are under-represented in a reference population, it would be best to supplement the phasing reference with as many ethnic-matched mutation carriers as possible.

Discussion

The traditional "family-based" haplotype phasing method is well established, but it has a number of shortcomings, such as lengthy preparation time and the need to recruit multiple family members. In contrast, new population-based phasing methods are unaffected by these issues. However, in the clinical setting, where the precision of pre-implantation test results cannot be compromised, it is crucial to implement the most accurate and reliable diagnostic methods available. The goal of this study was to determine whether population-based methods can achieve clinical-grade accuracy for PGD applications.

Recently, large-scale population-specific genomic databases were generated for various populations (including Icelandic [36], British [37], North American [38], Chinese [39], Korean [40], and Saudi Arabian [41] populations) mostly for enabling genome-wide association studies of rare variants. Other important applications of population-specific reference panels are in clinical genetics (look up of variant frequency), as well as in population genetics and evolutionary biology. Similar to other populations, a large-scale genomic data set also exists for Ashkenazi Jews [34, 35]. The extensive reference data can also be used to empower population-based phasing approaches, as we demonstrate here.

In order to demonstrate the feasibility of these ultrafast haplotype phasing strategies, we first tested whether a large cohort (1092 samples) from the 1000 Genomes Project (Caucasian but not Ashkenazi) would provide enough data to perform accurate phasing for our study cohorts. However, the poor accuracy obtained (Fig. 2) precluded the usage of the 1000G reference panel alone for haplotype phasing of the *CFTR* gene locus. When using Ashkenazi genomes as reference, we found, as expected, that the phasing accuracy increased with the size of the reference panel (Fig. 2). For individuals with mixed Ashkenazi and non-Ashkenazi ancestry, the accuracy was also higher when using the Ashkenazi-specific 574 samples reference panel compared to 1000G, but the overall accuracy of 95.4% ± 4.6 implicated the need for obtaining additional reference samples that are ethnically matched to the mixed ancestry population.

We next concentrated on founder mutations, since in these cases the haplotype carrying the mutation is expected to be relatively common. We chose cystic fibrosis for proof-of-principle, because it is a common disease (not only in the Ashkenazi population), and a test for *CFTR* mutations is included in all premarital/prenatal screening panels in Israel and elsewhere. We focused on the relatively common W1282X Ashkenazi-specific mutation and the delF508 mutation, which represents 70% of CFTR mutations in the non-Jewish European population [42, 43].

For patients of full Ashkenazi ancestry carrying the W1282X variant, using the 574 Ashkenazi genomes as a reference provided very high haplotyping accuracy. However,

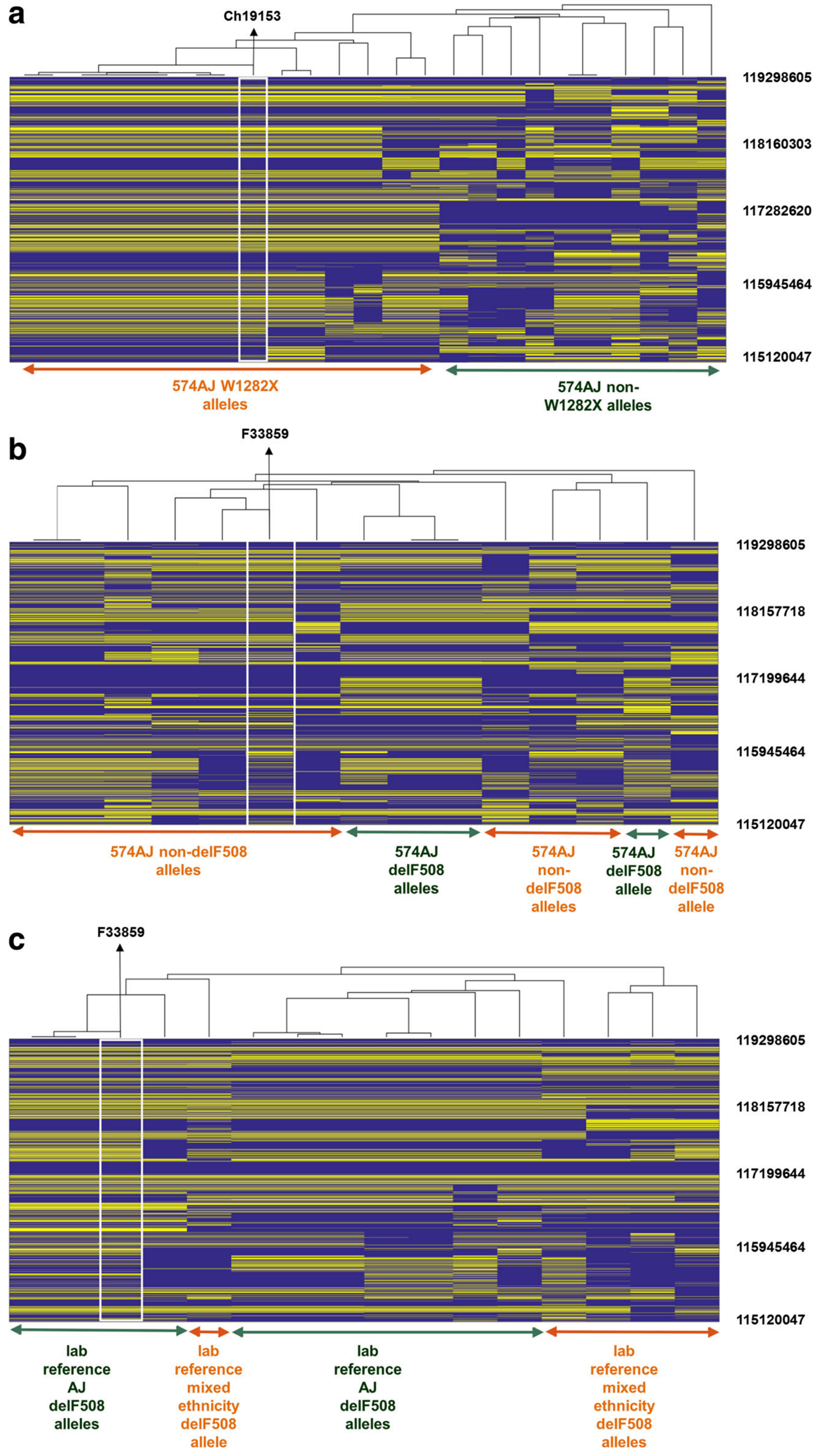


Fig. 4 Hierarchical clustering of W1282X and delF508 carriers in the study cohort after IBD-phasing. Pictured are the same dendrograms of Fig. 3, except that one unknown mutation carrier was added to each plot (enclosed by a white rectangle). **a** Sample Ch19153 (indicated on plot) carries the W1282X mutation in *CFTR*, yet was not accurately phased by short-range population-based modeling (SHAPEIT) using the 574 AJ reference panel (64.5% phasing accuracy; Table 2). However, after IBD-phasing with 574 AJ W1282X alleles, the phasing accuracy improved to 100% (Table 2). **b** Sample F33859 carries the delF508 mutation, and was poorly phased by SHAPEIT using the 574 AJ reference panel (68.9%; Table 2). IBD-phasing with delF508 carriers from the 574 AJ panel failed to improve accuracy (66.9%; Table 2). Note that sample F33859 incorrectly clustered with non-delF508 alleles. **c** The same delF08 lab-specific reference as that in Fig. 3c was used for IBD-phasing of sample F33859. Here, the delF08 allele of sample F33859 most closely matched the profile of one of the high quality trio-phased delF508 haplotypes in the lab reference. This improved phasing accuracy of F33859 from 68.9% (574AJ) to 96.4% (Table 2)

patients of mixed origin carrying the same mutation had a much lower phasing accuracy. To address this issue, we simply performed identity by descent analysis with W1282X alleles from the same reference data set to discriminate the W1282X-specific haplotype from the non-Ashkenazi haplotype in the "background."

For patients of full Ashkenazi ancestry carrying the delF508 mutation, not all individuals were accurately phased using the 574 AJ reference genomes, despite their full Ashkenazi ancestry. Given the high diversity of delF08 alleles in the AJ population (Fig. 3b), we developed a new reference panel of trio-based haplotypes from an in-house resource of

DNA from patients who have previously undergone PGD for delF508 in our clinic. These "in-house" reference samples turned out to capture much better the haplotype diversity in our study cohort and led to a significant increase in population-based phasing accuracy.

Thus, we show that population-based phasing is practical and accurate enough for clinical haplotype-based applications (such as PGD) according to the following guidelines. In general, individuals of FA descent are phased very accurately with a sufficiently large reference panel (such as 574AJ). The exception to this rule are FA carriers of rare, yet "diverse," variants (such as delF508) for which multiple founder alleles exist. For these exceptional cases, the base population reference should be supplemented by as many founder variant-bearing samples as possible to capture the full allelic representation of the variant of interest. When using this supplementation approach, population-based phasing features near perfect accuracy.

Regarding MA individuals, it would seem that 574AJ population-based phasing alone may not be reliable enough for clinical application. However, if the reference panel includes a sufficient number of carriers of a particular mutation with low allelic diversity (such as W1282X), IBD phasing is a rather simple yet attractive alternative to standard trio sample phasing. Within the AJ population, founder mutations are commonplace. Therefore, a computational solution to MA founder mutation carrier phasing (such as IBD) has practical utility for a myriad of pre-PGD case workups.

For populations other than Ashkenazi Jews, "hassle-free" accurate phasing can be performed based on publicly available data sets such as the 1000 Genomes Project [13] or the Haplotype Reference Consortium (HRC) [14]. The HRC contains over 32,000 whole genomes of individuals of various ancestries. In addition, dedicated servers exist (e.g., https://imputation.sanger.ac.uk/ or https://phasingserver.stats.ox.ac.uk/) for user-friendly phasing of target genomes based on the HRC and state-of-the-art phasing algorithms. Several other populations not represented in these data sets nevertheless have their own genome projects, such as Singaporeans [44], Mongolians [45], Costa Ricans [46], Qatari [47], French Canadians [48], and various Africans [49], to give a few examples. Even larger databases exist for microarray genotyping, such as the UK Biobank with nearly 500,000 individuals [50], although in that case, the imputed genotypes will have to be used to phase some of the SNPs.

In summary, we conclude that SNP data from existing whole genome sequences can be used to phase individuals of full AJ descent or partial AJ descent with founder mutation carriage. These findings essentially eliminate the need for family-based haplotyping and the extensive time required for associated "classical" family-based haplotype construction. Thus, it may soon be possible to employ clinical "OTSP," "Off-The-Street population-based Phasing," provided that one has access to an appropriate population-matched reference data set of sufficient size and diversity. Given the multitude of genome data that has and is currently being generated en masse for many worldwide populations, we predict that OTSP will soon be relevant for PGD seeking couples of any and all ethnicities in the near future.

Funding information The authors thank the Shaare Zedek Mirsky intramural grant and Rabbi David Fuld for funding this research. S. C. thanks the Israel Science Foundation grant 407/17.

Compliance with ethical standards

All individuals agreed to use their DNA samples for this study, and ethical approval was obtained according to Shaare Zedek Medical Center institutional review board guidelines.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Handyside AH, Kontogianni EH, Hardy K, Winston RM. Pregnancies from biopsied human preimplantation embryos sexed by Y-specific DNA amplification. Nature. 1990;344(6268):768–70. https://doi.org/10.1038/344768a0.
2. Yan L, Huang L, Xu L, Huang J, Ma F, Zhu X, et al. Live births after simultaneous avoidance of monogenic diseases and chromosome abnormality by next-generation sequencing with linkage analyses. Proc Natl Acad Sci U S A. 2015;112(52):15964–9. https://doi.org/10.1073/pnas.1523297113.
3. Thornhill AR, Handyside AH, Ottolini C, Natesan SA, Taylor J, Sage K, et al. Karyomapping-a comprehensive means of simultaneous monogenic and cytogenetic PGD: comparison with standard approaches in real time for Marfan syndrome. J Assist Reprod Genet. 2015;32(3):347–56. https://doi.org/10.1007/s10815-014-0405-y.
4. Ottolini CS, Rogers S, Sage K, Summers MC, Capalbo A, Griffin DK, et al. Karyomapping identifies second polar body DNA persisting to the blastocyst stage: implications for embryo biopsy. Reprod BioMed Online. 2015;31(6):776–82. https://doi.org/10.1016/j.rbmo.2015.07.005.
5. Natesan SA, Handyside AH, Thornhill AR, Ottolini CS, Sage K, Summers MC, et al. Live birth after PGD with confirmation by a comprehensive approach (karyomapping) for simultaneous detection of monogenic and chromosomal disorders. Reprod BioMed Online. 2014;29(5):600–5. https://doi.org/10.1016/j.rbmo.2014.07.007.
6. Natesan SA, Bladon AJ, Coskun S, Qubbaj W, Prates R, Munne S, et al. Genome-wide karyomapping accurately identifies the inheritance of single-gene defects in human preimplantation embryos in vitro. Genet Med. 2014;16(11):838–45. https://doi.org/10.1038/gim.2014.45.
7. Handyside AH, Harton GL, Mariani B, Thornhill AR, Affara N, Shaw MA, et al. Karyomapping: a universal method for genome wide analysis of genetic disease based on mapping crossovers between parental haplotypes. J Med Genet. 2010;47(10):651–8. https://doi.org/10.1136/jmg.2009.069971.
8. Handyside AH. Live births following karyomapping—a "key" milestone in the development of preimplantation genetic diagnosis. Reprod BioMed Online. 2015;31(3):307–8. https://doi.org/10.1016/j.rbmo.2015.07.003.

9. Gould RL, Griffin DK. Karyomapping and how is it improving preimplantation genetics? Expert Rev Mol Diagn. 2017;17(6):611–21. https://doi.org/10.1080/14737159.2017.1325736.
10. Dimitriadou E, Melotte C, Debrock S, Esteki MZ, Dierickx K, Voet T, et al. Principles guiding embryo selection following genome-wide haplotyping of preimplantation embryos. Hum Reprod. 2017;32(3):687–97. https://doi.org/10.1093/humrep/dex011.
11. Ben-Nagi J, Wells D, Doye K, Loutradi K, Exeter H, Drew E, et al. Karyomapping: a single centre's experience from application of methodology to ongoing pregnancy and live-birth rates. Reprod BioMed Online. 2017;35:264–71. https://doi.org/10.1016/j.rbmo.2017.06.004.
12. Zamani Esteki M, Dimitriadou E, Mateiu L, Melotte C, Van der Aa N, Kumar P, et al. Concurrent whole-genome haplotyping and copy-number profiling of single cells. Am J Hum Genet. 2015;96(6):894–912. https://doi.org/10.1016/j.ajhg.2015.04.011.
13. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. https://doi.org/10.1038/nature15393.
14. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48(10):1279–83. https://doi.org/10.1038/ng.3643.
15. Consortium UK, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526(7571):82–90. https://doi.org/10.1038/nature14962.
16. Loh PR, Palamara PF, Price AL. Fast and accurate long-range phasing in a UK Biobank cohort. Nat Genet. 2016;48(7):811–6. https://doi.org/10.1038/ng.3571.
17. O'Connell J, Sharp K, Shrine N, Wain L, Hall I, Tobin M, et al. Haplotype estimation for biobank-scale data sets. Nat Genet. 2016;48(7):817–20. https://doi.org/10.1038/ng.3583.
18. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48(11):1443–8. https://doi.org/10.1038/ng.3679.
19. Choi Y, Chan AP, Kirkness E, Telenti A, Schork NJ. Comparison of phasing strategies for whole human genomes. PLoS Genet. 2018;14(4):e1007308. https://doi.org/10.1371/journal.pgen.1007308.
20. Williams AL, Patterson N, Glessner J, Hakonarson H, Reich D. Phasing of many thousands of genotyped samples. Am J Hum Genet. 2012;91(2):238–51. https://doi.org/10.1016/j.ajhg.2012.06.013.
21. Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet. 2008;40(9):1068–75. https://doi.org/10.1038/ng.216.
22. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nat Rev Genet. 2011;12(10):703–14. https://doi.org/10.1038/nrg3054.
23. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084–97. https://doi.org/10.1086/521987.
24. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34(8):816–34. https://doi.org/10.1002/gepi.20533.
25. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6):e1000529. https://doi.org/10.1371/journal.pgen.1000529.
26. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. 2011;9(2):179–81. https://doi.org/10.1038/nmeth.1785.
27. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. 2013;10(1):5–6. https://doi.org/10.1038/nmeth.2307.
28. Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. Haplotype estimation using sequencing reads. Am J Hum Genet. 2013;93(4):687–96. https://doi.org/10.1016/j.ajhg.2013.09.002.
29. Daetwyler HD, Wiggans GR, Hayes BJ, Woolliams JA, Goddard ME. Imputation of missing genotypes from sparse to high density using long-range phasing. Genetics. 2011;189(1):317–27. https://doi.org/10.1534/genetics.111.128082.
30. Hickey JM, Kinghorn BP, Tier B, Wilson JF, Dunstan N, van der Werf JH. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genetics, selection, evolution : GSE. 2011;43(12):12. https://doi.org/10.1186/1297-9686-43-12.
31. Palin K, Campbell H, Wright AF, Wilson JF, Durbin R. Identity-by-descent-based phasing and imputation in founder populations using graphical models. Genet Epidemiol. 2011;35(8):853–60. https://doi.org/10.1002/gepi.20635.
32. SHAPEIT. https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html#home. Accessed July 19, 2018 2018.
33. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65. https://doi.org/10.1038/nature11632.
34. Carmi S, Hui KY, Kochav E, Liu X, Xue J, Grady F, et al. Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. Nat Commun. 2014;5:4835. https://doi.org/10.1038/ncomms5835.
35. Lencz T, Yu J, Palmer C, Carmi S, Ben-Avraham D, Barzilai N, et al. High-depth whole genome sequencing of an Ashkenazi Jewish reference panel: enhancing sensitivity, accuracy, and imputation. Hum Genet. 2018;137(4):343–55. https://doi.org/10.1007/s00439-018-1886-z.
36. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015;47(5):435–44. https://doi.org/10.1038/ng.3247.
37. Genome England. http://genomicsengland.co.uk. Accessed 2018 2018.
38. All of Us. https://allofus.nih.gov/. 2018.
39. Cyranoski D. China embraces precision medicine on a massive scale. Nature. 2016;529(7584):9–10. https://doi.org/10.1038/529009a.
40. Korean Reference Genome Project. http://152.99.75.168/KRGDB/menuPages/intro.jsp. 2018.
41. Abu-Elmagd M, Assidi M, Schulten HJ, Dallol A, Pushparaj P, Ahmed F, et al. Individualized medicine enabled by genomics in Saudi Arabia. BMC Med Genet. 2015;8(Suppl 1):S3. https://doi.org/10.1186/1755-8794-8-S1-S3.
42. Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, et al. Identification of the cystic fibrosis gene: genetic analysis. Science. 1989;245(4922):1073–80.
43. Worldwide survey of the delta F508 mutation—report from the cystic fibrosis genetic analysis consortium. Am J Hum Genet. 1990;47(2):354–9.
44. Wu D, Dou J, Chai X, Bellis C, Wilm A, Shih CC, et al. Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. bioRxiv. 2018. https://doi.org/10.1101/390070.
45. Bai H, Guo X, Narisu N, Lan T, Wu Q, Xing Y, et al. Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and

East Asia. Nat Genet. 2018;50:1696–704. https://doi.org/10.1038/s41588-018-0250-5.

46. Mooney JA, Huber CD, Service S, Sul JH, Marsden CD, Zhang Z, et al. Understanding the hidden complexity of Latin American population isolates. Am J Hum Genet. 2018;103(5):707–26. https://doi.org/10.1016/j.ajhg.2018.09.013.

47. Fakhro KA, Staudt MR, Ramstetter MD, Robay A, Malek JA, Badii R, et al. The Qatar genome: a population-specific tool for precision medicine in the Middle East. Human Genome Variation. 2016;3:16016. https://doi.org/10.1038/hgv.2016.16.

48. Low-Kam C, Rhainds D, Lo KS, Provost S, Mongrain I, Dubois A, et al. Whole-genome sequencing in French Canadians from Quebec. Hum Genet. 2016;135(11):1213–21. https://doi.org/10.1007/s00439-016-1702-6.

49. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015;517(7534):327–32. https://doi.org/10.1038/nature13997.

50. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203–9. https://doi.org/10.1038/s41586-018-0579-z.