Detection of sharing by descent, long-range phasing and haplotype imputation

Augustine Kong[1], Gisli Masson[1], Michael L. Frigge[1], Arnaldur Gylfason[1], Pasha Zusmanovich[1], Gudmar Thorleifsson[1], Pall I. Olason[1], Andres Ingason[1], Stacy Steinberg[1], Thorunn Rafnar[1], Patrick Sulem[1], Magali Mouy[1], Frosti Jonsson[1], Unnur Thorsteinsdottir[1], Daniel F. Gudbjartsson[1], Hreinn Stefansson[1], Kari Stefansson[1].

Corresponding authors:
Augustine Kong, deCODE genetics, Sturlugata 8, 101 Reykjavik, Iceland.
kong@decode.is, Phone: 354-5701931, fax 354-5702850.
Kari Stefansson, deCODE genetics, Sturlugata 8, 101 Reykjavík, Iceland.
kari.stefansson@decode.is, Phone:354-5701900, fax 354-5701901.

[1]deCODE genetics, Sturlugata 8, 101 Reykjavík, Iceland

Supplementary Material

Supplementary Material is divided into three sections, Supplementary Methods, Supplementary Results, and Supplementary Note.

1. Supplementary Methods

1.1 Algorithm used to resolve incompatibilities when applying LRP

Incompatibilities in phase information were resolved as follows: First, the phase information provided by the surrogate parents was used to determine a putative consensus phase for the proband, using a simple majority rule at each marker. Markers for which there was no majority were declared unphased. Missing genotypes in surrogate parents did not contribute to the vote, while missing genotypes in the proband were not imputed. Then if the surrogate parents matching this putative consensus accounted for over 90% of the total phase information provided or if at most two surrogate parents, while still in minority, did not match the putative consensus, it was used as consensus for the proband. If not, the marker with the most discrepant phase information was declared unphased and the method was applied again to the all the data with this marker excluded. If consensus could not be reached in this way after removing 10% of the heterozygous markers for the proband, the proband was declared unphased on all markers. After Round 1, the phasing results at that point were used to prune the surrogate parent list as follows. For each proband A, every surrogate parent B that did not match the consensus phase for A was removed from the surrogate parent list of A. In order to maintain symmetry in the resulting surrogate parent list, A was also removed from the surrogate parent list of B. The algorithm was then run again using the original genotypes and the revised list of surrogate parents.

1.2 Formula for computing the Legacy Coefficient

The legacy coefficient is defined as the probability that a haplotype of a proband would be inherited by at least one child or grandchild who is typed. Note that for autosomes, the probability is the same for the paternal and maternal chromosomes of a proband. Let *M* be the total number of children of a proband, i.e. it includes both typed and untyped children. Let *K* be the number of untyped children, and for $i = 1, \ldots, K$, let n_i be the number of typed children (i.e. grandchildren of the proband) these untyped children have. The legacy coefficient can then be computed as

$$1 - \frac{1}{2^M} \prod_{i=1}^{K} \left(1 + \frac{1}{2^{n_i}}\right)$$

2. Supplementary Results

2.1 Phasing

The two regions on chromosome 15 phased in addition to the MHC region

Both regions include rs1051730 which is also mentioned in the haplotype imputation example in the main text. The long region was chosen to match the phased MHC region in physical length (10 Mb). The target region includes 895 SNPs and is defined by rs1564492 and rs7162082 (~ 5Mb respectively on the left and right of rs1051730). The region used to select putative surrogate parents is extended by ~ 2Mb on both sides, and includes a total of 1231 SNPs. The shorter region was designed to be similar to the phased MHC region in genetic length. The target region includes 574 SNPs and is defined by rs4886630 and rs3743421 (~ 3cM respectively on the left and right of rs1051730). The region used to select putative surrogate parents is extended by ~ 2cM on both sides, and includes a total of 979 SNPs.

Results for the Trio Test

Supplementary Table 1a. Comparing LRP without parents to trio phasing.

Discrepancies	MHC (N=2518)	C15 Long (N=2518)	C15 Short (N=2562)
0	2456	2459	2541
1	43	41	15
2	8	0	0
3	1	1	1
>3	10	17	5
Total	845/978802 (0.086%)	496/622148 (0.080%)	91/412361 (0.022%)

N = number of offspring/probands phased by both LRP and the standard trio method. The cells provide a summary of the differences observed between LRP performed without parents and the trio method. For example, for the MHC region, there are no discrepancies between the method for 2456 probands, and exactly one mismatch was observed for 43 probands. Out of the 978,802 heterozygous genotypes phased by both methods, a total of 845 mismatches, or 0.086%, were observed.

To investigate the contribution of typed siblings to phasing, we tabulated the results for the 1249 (out of 2718) offspring with no sibling typed. For the MHC, C15 Long, and C15 short, regions respectively, 89.2%, 90.2%, 92.5%, of the heterozygous SNPs were phased. As expected, these yields are lower than that for the 2718 offspring as a whole (Table 2), but the differences, ranging from 1.1% to 2.2%, are not substantial. The discrepancies compared to trio phasing are summarized in Supplementary Table 1b. When compared to results in Supplementary Table 1a, we see that the discrepancy rate is somewhat lower for the MHC region, essentially the same for the C15 Long region. It is higher for the C15 Short region, and it appears that most of the discrepancies for the 2718 offspring occurred here. But the discrepancy rate remains a very low 0.045%. Overall, we see that having siblings typed is certainly a plus, but it is not a major factor to the performance of LRP.

Supplementary Table 1b. Discrepancy results for the 1249 offspring with no sibling typed.

Discrepancies	MHC (N=1130)	C15 Long (N=1139)	C15 Short (N=1161)
0	1101	1110	1144
1	21	22	12
2	4	0	0
3	0	1	0
>3	4	6	5
Total	227/438312 (0.052%)	226/282365 (0.080%)	85/187841 (0.045%)

Comparisons with PHASE and FASTPHASE

Speed comparisons were made based on a 3.2 GHz Intel machine running linux. For the C15 short region, we started a run with PHASE for 300 individuals, and it did not finish after 3 weeks. For the same region, we were able to finish a run with fastPHASE processing 10,000 typed individuals. These individuals include the 2718 offspring in the trio test and another 7282 individuals who were randomly selected from the set of typed individuals with the parents in the trios removed. Performance of fastPHASE is compared to that of LRP applied to the same set of 10,000 individuals (LRP-10000), and LRP applied to 31,702 individuals (with the parents in the trios removed from the full set of 35,528) (LRP-31702). The run times were approximately 92 hours for fastPHASE, 16 minutes for LRP-10000, and 90 minutes for LRP-31702.

Two out of the 2718 offspring are homozygous for all 574 SNPs in this region, and for the purpose of comparing phasing accuracy, only the 2716 offspring who have heterozygous genotypes are considered here. For fastPHASE, LRP-10000, and LRP-31702, the proportion of heterozygous genotypes phased are 100%, 85.4% and 93.6% respectively. The discrepancy rate when compared to trio phasing, as illustrated in Supplementary Table 2, is 30.411%, 0.066% and 0.022% respectively. Compared to LRP-31702, LRP-10000 has lower yield, as expected, and a higher discrepancy rate. But considering that 10,000 individuals are only about 3% of the total living population in Iceland, we found these numbers encouraging.

The phasing results from fastPHASE and LRP do not contain information on origin, i.e. which haplotype is paternal and which is maternal. Hence there are two possible ways to compare the results with that of trio phasing. The discrepancy count is taken as the minimum of the two possible comparisons, and hence the discrepancy rate by definition cannot be higher than 50% for any individual. This means that a discrepancy rate of 30.411% for fastPHASE is really very high. Essentially, the phasing of SNPs that are separated by many LD blocks by fastPHASE, and probably for all other local phasing methods, is close to random.

Supplementary Table 2. Discrepancy rate comparisons for three ways of phasing

Discrepancies	fastPHASE (N=2716)	LRP-10000 (N=2390)	LRP-31702 (N = 2560)
0	17	2355	2539
1	11	21	15
2	6	1	0
3	16	0	1
>3	2666	13	5
Total	134008/440663 (30.411%)	248/376221 (0.066%)	91/412361 (0.022%)

Another useful way to compare the accuracy of the different methods is to look at the number of individuals who are correctly phased for the entire region. As noted earlier, a single discrepancy with trio phasing does not necessarily correspond to an error, as it could easily be a result of a genotyping error in one of the parents. Suppose we consider individuals with 2 or less discrepancies to be essentially correct all through. Based on this criterion, fastPHASE is correct for 1.25% (34/2716) of the individuals it phased, and the corresponding numbers are 99.46% (2376/2390) and 99.77% (2554/2560) for LRP-10000 and LRP-31702 respectively. Notice that log(0.0125)/log(0.9946) ~ 800, or $(0.9946)^{800}$ ~ 0.0125. For local phasing methods such as fastPHASE, the chance that a region is phased without error should decrease exponentially as a function of the length of the region. This implies that LRP-10000 can phase a region approximately 800 times longer than fastPHASE with a similar chance of not making any errors. Since log(0.0125)/log(0.9977) ~ 1800, this suggests that LRP-31702 can phase a region 1800 longer than fastPHASE with a similar chance of making any errors. We do however recognize that the very low error rate of LRP-31702 might partly be due to chance. Also, even though we do not think that the accuracy of fastPHASE could improve much by further increasing the sample size, comparing the accuracy of the current fastPHASE run, which utilized data from 10,000 individuals, with that of LRP-31702 might not be completely fair. Overall, we feel that it is a reasonable estimate that LRP is about 800 times as accurate as fastPHASE for our data. We note that similar conclusions would be reached if one or less discrepancy, as opposed to 2 or less, is used as the criterion for determining that phasing is essentially correct for the entire region.

We note that the accuracy estimate we arrived for fastPHASE is not inconsistent with what had been documented. According to Marchini *et al.* (ref 4 in the main text) which studied many of the local phasing methods, for unrelated individuals, the chance that a 1 Mb region, which is rather short by the standards here, can be correctly phased in its entirety is not very high.

2.2 Studying the recurrent deletion at 15q11.2.

For phasing, we started by employing the first 400 SNPs we had on chromosome 15. They span approximately 6.0 Mb (Build36: 18.4Mb to 24.4Mb, starting with rs6599770 and ending at rs1863459) and 8.0 cM. These include the 51 SNPs in the deletion region. We performed LRP in two ways: (1) with and (2) without the 51 SNPs. Note that those who were identified as surrogate parents of a proband with the deletion in both (1) and (2) are consistent with sharing the haplotype *without* the deletion, while those who were identified as surrogate parents in (2) but not in (1) are consistent with sharing the haplotype *with* the deletion. Note that both types of surrogate parents contribute to phasing the proband and in determining the haplotype background around the deleted region. However, those identified to carry the haplotype associated with the deletion, but not having the deletion themselves, further allowed us to determine the haplotype of the deleted SNPs. Moreover, with the assistance of the genealogy, as demonstrated in the example, we could sometimes determine the likely point in time at which the original mutation occurred.

After phasing, the haplotypes of each of the 63 chromosomes with the deletion was compared to the others. Those who were identified as surrogate parents of each other were established as IBD already, and we further determined the extent of sharing by going beyond the original 400 SNPs examined. However, those who were not identified as surrogate parents of each other through the phasing procedure could still be IBD, i.e. the shared region might be less than the 6.0 Mb that we examined. Relative to the IBS ≥ 1 criterion, comparison of two phased haplotypes for compatibility is substantially more specific, providing more power to distinguish true sharing from noise. When evaluating the statistical significance of the number of consecutive SNPs for which two haplotypes associated with the deletion were in agreement, the over 70,000 haplotypes on chromosomes without the deletion were used as the baseline. Apart from a couple of exceptions, deletions on two chromosomes were clearly IBD if the carriers were separated by 7 meioses or less, and not IBD otherwise. One exception was most likely a consequence of an inaccuracy in the genealogy; one carrier shares a long stretch of SNPs with two other closely related carriers, but was separated by 20 meioses from them according to the genealogy. It is possible that this proband was adopted. Another exception is with P2 in the example and a niece of hers who is also a carrier. Their haplotypes only agree for 47 SNPs. This is however nominally significant. Moreover, no other surrogate parent of the niece sharing the haplotype on the deleted chromosome was found. Hence, we believe that there is real chance that the deletions in the niece and P2 are IBD and the sharing was cut short by a recent recombination event nearby, but it is not definitive.

One might ask how much of the mutation history could be inferred from the pedigree alone without long range phasing. The answer is that the pedigree information certainly can help (e.g. the example of the niece discussed above), but it is far from adequate. The transmission of the deletion from parent to offspring when both are typed for the deletion is usually clear. When the parents of a carrier are genotyped but do not have the deletion, we can conclude that the deletion is *de novo* (one out of the 63 cases in this case). But when the parent who transmitted the chromosome with the deletion to the proband is not genotyped, based on the pedigree alone, we cannot know whether or not the deletion is *de novo*, nor can we decide whether or not two probands falling into this category have deletions that are IBD. Indeed, before applying LRP, some of us thought some carriers separated by 8 meioses might have deletions that are IBD, but the phasing results prove otherwise. Also, as noted above, the pedigree information is not perfect. Two carriers that appeared to be very distantly related might in effect be closely related.

3. Supplementary Note

3.1 Ways to improve the phasing algorithm in the future

1. Model-based rather than rule-based. The current algorithm for phasing and IBD detection is rule-based. While rule-based methods are usually easier to implement and often computationally less demanding, they are ultimately limited. Our goal is to develop a method that directly models recombination and genotyping error rate, two main factors that impact IBS/IBD sharing. A model-based method would also allow us to use the population frequencies of alleles and haplotypes to determine the likelihood that two individuals share a haplotype IBD. This is particularly important for situations where the shared haplotype does not involve a very large number of SNPs.
2. Considered as surrogate parents also those who do not have IBS ≥ 1 for the entire target/extended region. A. Incorporating people who only share a haplotype with the proband for part of the target/extended region. B. Allowing for the possibility that IBS = 0 for a single SNP in the target region might be a result of a genotyping error in the parent or the potential surrogate parent. In the case of A, sometimes a relative can share a very long haplotype with the proband, extending far outside the target region on one side, but the shared haplotype does not cover the entire target region. This relative can be used to phase the proband for part of the target region. Note that having one such relative each for the left and right sides of the target region can result in the phasing of the entire region. Doing B in a model-based manner becomes even more important when higher density chips are used (e.g. 1 million instead of 300,000 SNPs). Avoiding that can lead to the paradoxical situation of doing less with more data.
3. Employing the principles of local phasing. First, it is important to note that local phasing and LRP utilize data in different ways, and hence, in theory, one can always do better by employing both techniques simultaneously. Obviously, individuals who cannot be phased by LRP could be partially phased by local phasing. But it goes much further than that. Local phasing within solid LD blocks could in effect create markers that are much more polymorphic than a SNP. IBS sharing that is determined

using these more polymorphic markers has much more specificity. This can help in both filtering out false surrogate parents and better identification of true surrogate parents that share a shorter region. Also, local phasing can assist in resolving discrepancies. When LRP, as implemented currently, could not reliably resolve which of two phasing possibilities is correct (usually corresponds to having to discard one of two groups of putative surrogate parents which are incompatible with each other), local phasing can be used to evaluate which of the two possibilities is more likely.

4. Distinguish first degree relatives ('real' instead of surrogate, i.e. parent-offspring and full-sibs) from other relatives at the phasing stage. Note that these relations can often be established by the genome-wide data without the genealogy. Doing this can enhance phasing in two ways. (A) Avoid large phasing errors (i.e. those that affect a large number of SNPs). Close relatives, who can share two different haplotypes with the proband, sometimes at different loci and sometimes at the same loci (with full-sibs) can create confusion in phasing and lead to large errors. Knowing who they are makes it easier to avoid committing such mistakes. (B) Identifying genotyping errors. This can help both to increase yield (e.g. a real parent would not be ignored as a surrogate parent because of IBS = 0 for one SNP) and also reduce phasing errors.

3.2 Applying LRP to other datasets

Here we list what we consider might be key to the application of LRP to data other than those we studied in Iceland. This is not meant to be an exact roadmap, but rather an attempt to document what we have learned from our experience.

1. Empirical Quality Control and Parameter Adjustment using Trios. As part of the typed sample, having a substantial number of trios is important. As demonstrated, the trios can be used to empirically evaluate the accuracy of a method. Most importantly, the trios allow us to compare different methods and to choose the optimal parameter settings. Although more is always better, having 100 trios may be enough. That is because the genome can be partitioned into many different regions and each can provide a test/comparison that is close to independent of the others.

2. Sample Size, Population Structure and Yield. As noted (Table 1 and Figure 2), the yield of LRP is directly tied to the expected number of surrogate fathers and surrogate mothers that the typed individuals have, which is a function of the sample size and the average kinship coefficient among the typed individuals. Hence if an estimate of average kinship coefficient is available for a population, it can be used to predict the yield of LRP as a function of sample size. Note that the kinship coefficient between two individuals depends on how many generations up we go. In Table 1, the calculation is based on the pedigree of Iceland going back to 1650, or about 10 generations ago. The average kinship coefficient will increase if we go further back in time. If two individuals share a region IBD inherited from a common ancestor 10 generations ago, the length of the shared region is on average 10 cM. But there is substantial variation, e.g. it can easily be twice as long or only half as long. While this means that some IBD sharing that are separated by less than 20 meioses would not be captured by LRP, this also means that some that are separated by more than 20 meioses would. If an estimate of average kinship coefficient is not available for a population, we believe that the results in Figure 2, with the focus on the fraction of the living population genotyped, as opposed to the absolute number, could be a useful guide. The reasoning is as follows. Consider a population that has a similar fertility rate as Iceland in the last 10 generations and suppose that the population is reasonably closed in the sense that most of the relatives of an individual in the population, up to 10 generations, are also in the population. Under these assumptions, (A) the number of cousins, second-cousins, etc. of an individual in this other population will be similar to an individual in Iceland, and (B) for the same fraction of the population genotyped, the chance that a particular cousin, or second cousin, etc, would be genotyped is the same. With (A) and (B), the yield of LRP should be similar for the two populations. Suppose this other population is very large, one might notice that a person there might on average have, for example, more fourth-cousins than a person in Iceland. Even though not inbred, Iceland is a small population, and many individuals are related through multiple paths in the pedigree, hence two individuals can easily be four-cousins two ways, and hence the number of distinct individuals who are four-cousins would be reduced. But this complication is really not an issue

since the counting in (A), for the sake of LRP, should be based on relationship paths instead of distinct individuals, with one exception. Two individuals sharing both haplotypes IBD is not useful for phasing and, in a sense, the IBD sharing is 'wasted'. In some instances, as noted, the sharing of both haplotypes at some locus might even lead to phasing errors. Even though the effect appears to be small, the chance of that happening is higher in Iceland than in a larger population. Also, this could be a substantial issue in a small inbred population. Hence, one might conclude that, for the same fraction of the population genotyped, LRP may actually work better in a larger population than a smaller one. It is also noted that for populations primarily made up of recent immigrants and their descendants, a larger fraction of the population would need to be typed to achieve comparable results.

3. Length of the Phased Region and Yield. Obviously more people share a shorter region than a longer one. This is directly related to the point noted above that the average kinship coefficient increases as we reach further back in time. When viewing Figures 2a and 2b, the focus should not be on the difference of yield for the same sample size, but rather, as noted in the text, relative to the MHC and C15 Long regions, similar yield was attained for the C15 short region with about ¾ (1.5% versus 2% of the population phased) the sample size. Notice that the C15 Short region is still 6 cM (6.4 Mb) in length. For example, phasing a 3 cM region accurately, which could be extremely useful for many applications, should be achievable with an even smaller sample size.
4. Genotype Quality. Genotype quality is extremely important. For LRP, the demand for genotype accuracy is even higher than that for association studies. We estimated our genotype error rate to be about 0.01%. If the error rate is 0.1% or 1%, our current algorithm can end up having much lower yield. Future improvements to the algorithms made by us and others are expected to adapt to the specific genotyping error rate of the data and become more tolerant. Still, even sophisticated statistical modeling cannot be expected to fully compensate for low quality genotypes. Hence, from the design view point, for those who are planning for future genotyping and have interest in applying LRP, this should be a very important consideration. For existing genotypes, there are a number of factors to consider. Error rate varies

between SNPs, and it is probably better to err on the conservative side by not using SNPs that could have a higher error rate than average. If these SNPs are important, they could in theory be added back in after the initial phasing is done. The latter step could also utilize local phasing information.

5. SNP density. In theory, with more SNPs and hence more information, the performance of LRP phasing should improve. For example, for a fixed region, with more SNPs, the chance for two individuals to have IBS ≥ 1 for the whole region without sharing a haplotype IBD will be reduced. This can increase yield since the IBS ≥ 1 criterion can then be applied to a shorter region when identifying putative surrogate parent. However, to capitalize on this increase of information, the current algorithm must be improved to better handle genotyping errors. If not, yield can actually decrease because of the increased number of discrepancies resulting from genotyping errors.

Finally, as others have also noted that IBD sharing can often be detected using high density SNP data without direct pedigree information, (e.g. Nelson, S. *et al*. Detecting identical-by-descent DNA intervals between affected distant relatives using high-density SNP genotyping., oral presentation 44, Annual ASHG meeting, Salt Lake City, Utah), there is no reason to believe that LRP cannot be applied to other datasets under the right conditions.