A. Supplementary materials

A.1. The algorithm of Anchap

Algorithm 1 Stage I: first scan for sharing from unphased genotypes

{Input: multi-locus genotypes for individuals in the cohort}
for all individuals i in the cohort **do**
    $i.g \Leftarrow i^{th}$ genotype across all markers
    for all individuals j in the cohort, such that $i \neq j$ **do**
        $j.g \Leftarrow j^{th}$ genotype across all markers
        between $i.g$ and $j.g$ find regions (start and end) without opposing homozygotes longer than the IBD threshold
    end for
end for
{Output: list of genomic regions and pairs of individuals putatively sharing IBD}

Algorithm 2 Stage II: alignment of haplotypes

{Input: multi-locus genotypes for individuals in the cohort, list of genomic regions and pairs of individuals putatively sharing IBD}
for all individuals i in the cohort **do**
  sort the sequences shared with i by the number of markers they cover, descending
  for all s, shared sequences of i **do**
    $i.hap.pat.s \Leftarrow$ current version of i^{th} paternal haplotype, in the region of sharing s
    $i.hap.mat.s \Leftarrow$ current version of i^{th} maternal haplotype, in the region of sharing s
    $s.g \Leftarrow$ the genotype of the individual sharing with i in region s
    if $s.g$ is matching $i.hap.pat.s$ and $i.hap.mat.s$ and no other sequences have been seen in the region before **then**
      $s.g$ shares i's paternal haplotype (arbitrarily)
    else if $s.g$ is matching $i.hap.pat.s$ **then**
      $s.g$ shares i's paternal haplotype
    else if $s.g$ is matching $i.hap.mat.s$ **then**
      $s.g$ shares i's maternal haplotype
    else
      reject $s.g$
    end if
    use $s.g$ to recover the relevant haplotype
  end for
end for
{Output: revised list of genomic regions and pairs of individuals putatively sharing IBD, with each region assigned to individuals' haplotypes, genotype phasing in IBD regions}

Algorithm 3 Stage III: Second scan for sharing from partially complete haplotypes

1: {Input: genotypes phased in IBD regions}
2: **for all** individuals i in the cohort **do**
3: $i.hap.pat \Leftarrow i^{th}$ paternal haplotype, from round 1
4: $i.hap.mat \Leftarrow i^{th}$ maternal haplotype, from round 1
5: **for all** individuals j in the cohort, such that $i \neq j$ **do**
6: $j.hap.pat \Leftarrow j^{th}$ paternal haplotype, from round 1
7: $j.hap.mat \Leftarrow j^{th}$ maternal haplotype, from round 1
8: between haplotypes of i and j, find continuously matching regions longer than the IBD threshold AND with at least this many markers where alleles on both haplotypes are fully known
9: **end for**
10: **end for**
11: {Output: revised list of genomic regions and pairs of individuals putatively sharing IBD, with each region assigned to individuals' haplotypes, genotype phasing in IBD regions}

Algorithm 4 Choosing an optimal subset of individuals for re-sequencing

{Input: revised list of genomic regions and pairs of individuals putatively sharing IBD}
cov.mat $\Leftarrow$ a matrix of size number of haplotypes (2x number of individuals) by number of markers, to store whether, with individuals picked so far, each locus of a haplotype has an individual sequenced who shares this haplotype at this region; initialized to 0 everywhere
picked.inds $\Leftarrow$ and empty list of individuals to be picked for the study in the preferred order
while not chosen the desired number of individuals **do**
  for all individuals i in the cohort, s.t. i not in *picked.inds* **do**
    $i.cov \Leftarrow 0$
    for all individuals j in the cohort, s.t. j not in *picked.inds* **do**
      $i.cov.j \Leftarrow$ total length of shared haplotype segments between i and j, such that *cov.mat* in the corresponding regions is 0
      $i.cov \Leftarrow i.cov + i.cov.j$
    end for
  end for
  for sequencing choose next individual i, s.t. $i = max_i i.cov$ and i not in *picked.inds*
  add i to *picked.inds*
  for all haplotypes hj in the cohort **do**
    for all IBD regions r between i and hj **do**
      $cov.mat.hj.r \Leftarrow 1$
    end for
  end for
end while
{Ouput: preferred sequencing order}

A.2. Example of ANCHAP

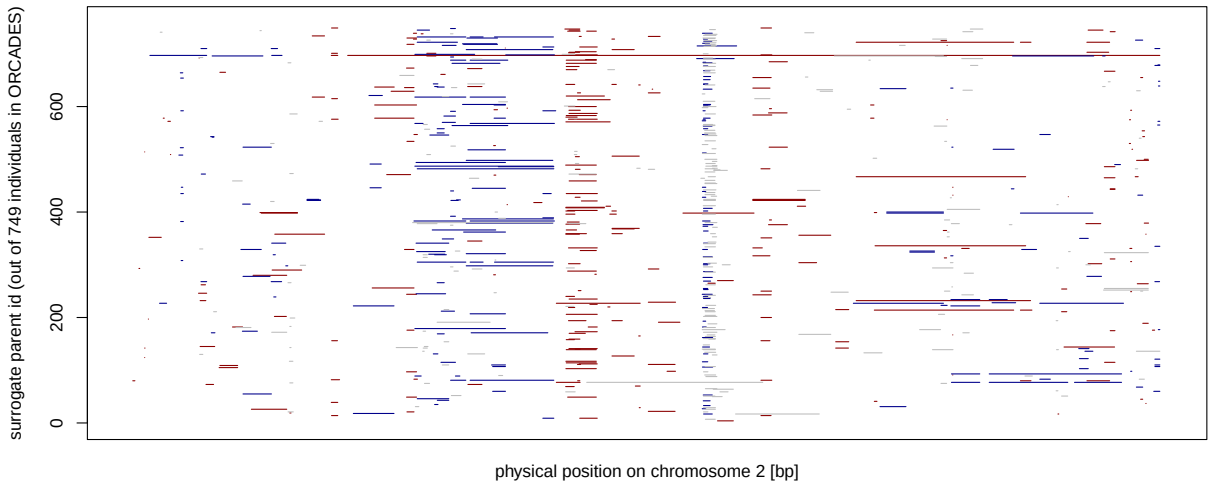


Fig. 1.—: Example of IBD detection (Stage I), alignment of IBD regions (Stage II) and phasing (Stage III) of one individual from ORCADES, chromosome 2. First we find his/her haplotype sharers across the genome, and mark the regions of putative IBD sharing as segments. The shared sequences are aligned into two groups, and marked red, and blue accordingly. The grey segments denote misaligned shared sequences. Individual 697 is a full sibling of the proband, with almost the entire chromosome shared and more distant relatives share smaller blocks.

A.3. Data pre-processing

The data sets were pre-processed in PLINK (Purcell et al. 2007) to eliminate low quality markers. We removed markers with call rate of less than 95%, out of Hardy-Weinberg equilibrium ($p < 0.001$), or those with minor allele frequency lower than 1%. We excluded individuals with more than 7% genotype markers missing, and retained only the autosomal SNPs. After pre-processing, the following numbers of samples remained: ORCADES (749 individuals, 302,379 SNPs on 22 chromosomes), CROATIA-KORCULA (945 individuals; 317,223 autosomal SNPs, including 295,574 ORCADES SNPs), CROATIA-VIS (991 individuals; 301,069 autosomal SNPs, including 291,857 ORCADES SNPs), SOCCS (958 individuals; 306,204 autosomal SNPs, including 294,703 ORCADES SNPs). We localized the SNPs on the HapMap genetic map of recombination rates (Consortium 2007).

A.4. Regions of increased frequency of IBD

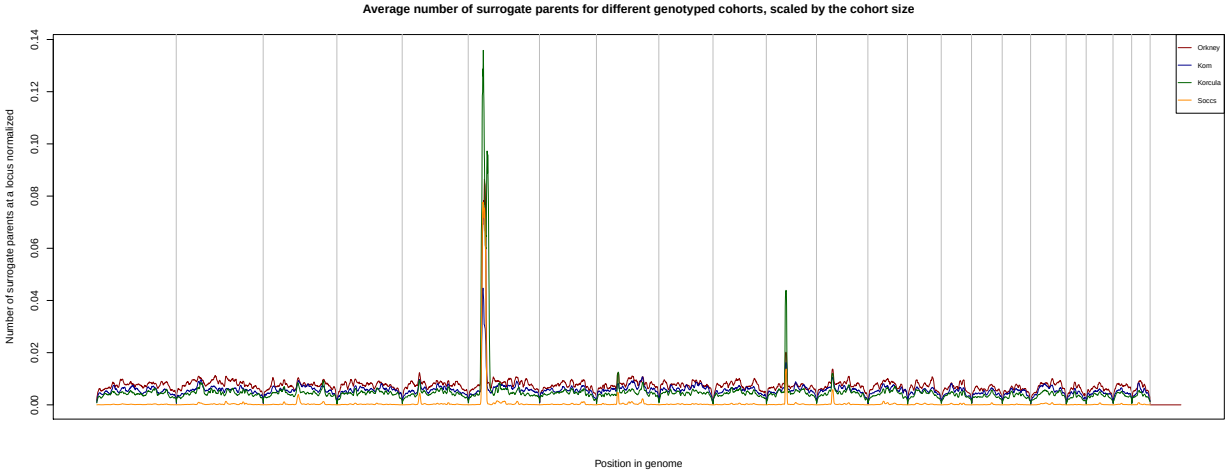


Fig. 2.—: IBD peaks on chromosomes 6 and 11, before a genetic map was used to account for extensive LD amoung the isolate founders. The peak on chromosome 6 was reduced and the one on chromosome 11 almost compeltely removed, when we used the HapMap genetic map.

Table 3 shows genetic positions of the peaks of IBD, which were marked at the horizontal axes in Figure 2 of the main article.

chromosome	position left [kb]	position right [kb]
2	134144	138947
3	15484	24365
6	27145	33161
8	95306	97626
10	100639	119196
14	77965	88690
19	18379	34464

(a) peaks in IBD density for ORCADES (build 36)

chromosome	position left [kb]	position right [kb]
1	185353	189270
2	47210	59831
6	25952	33936
9	78602	81335
9	99130	104709

(b) peaks in IBD density for CROATIA-VIS (build 35)

chromosome	position left [kb]	position right [kb]
1	90094	101013
1	167719	177243
2	54456	63368
12	77079	90001
18	64446	66196

(c) peaks in IBD density for CROATIA-KORCULA (build 36)

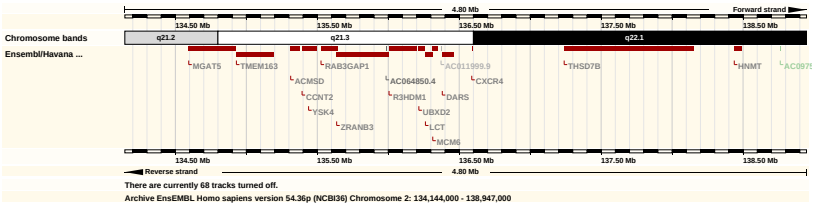
chromosome	position left [kb]	position right [kb]
2	134028	139092
6	25535	33096

(d) peaks in IBD density for SOCCS (build 36)

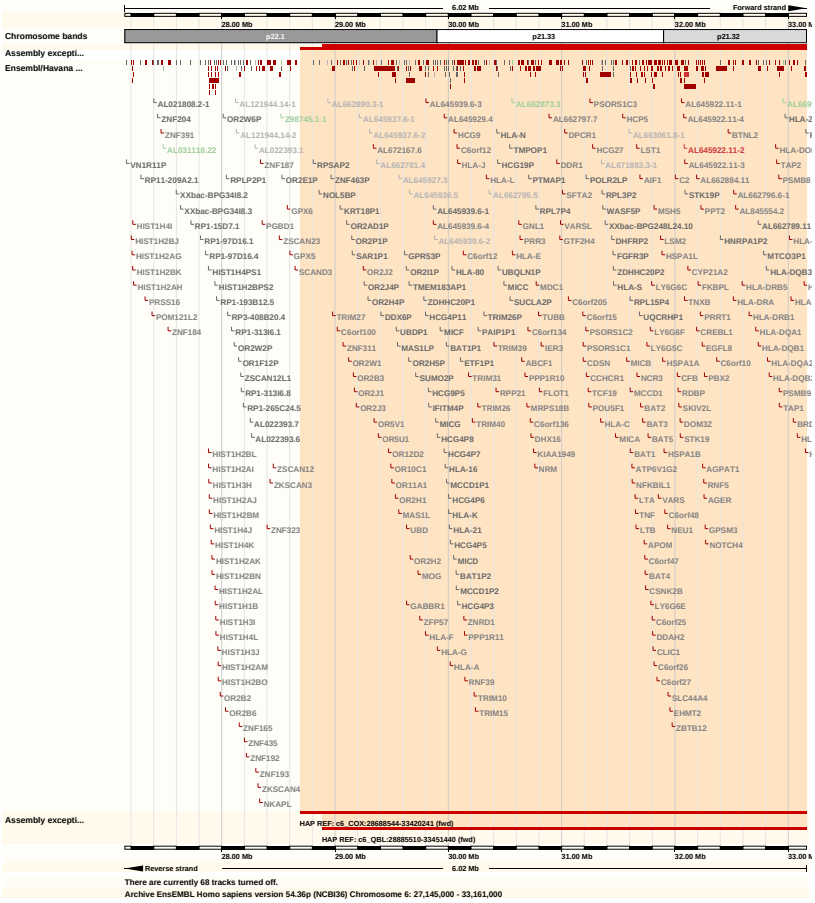
Fig. 3.—: Locations of IBD peaks in four cohorts under study

A.5. Peaks of IBD - interpretation

Figure 4 shows genes present in the two top peaks on chromosomes 2 and 6.



(a) peak on chromosome 2, between 134144 and 138947 kb



(b) peak on chromosome 6, between 27145 and 33161 kb

Fig. 4.—: Genes in the regions of increased IBD

A.6. Parameter tuning and performance metrics

Below we describe tuning of the parameters of ANCHAP, which include:

- IBD threshold (Stage I)
- IBD region margins (Stage I)
- alignment parameters: overlap threshold and matching threshold (Stage II)
- number of markers phased for both individuals in a putative IBD region (Stage III)

Tuning is informed by the following performance metrics:

- evaluation against reference recent IBD - the results between the reference individuals were evaluated against the regions of true recent IBD. The total number of markers in true regions and in resulting regions is TP, in true regions but not in the resulting regions is FN, not in true regions but in the resulting regions FP, and neither in the true regions nor in the resulting regions TN. From these, sensitivity and false discovery rate can be computed.
- sensitivity - $TP/(TP + FN)$
- false discovery rate - $FP/(FP + TP)$
- inconsistency rate - how many of the alleles of haplotype sharers were homozygotes not consistent with homozygotes of the majority of the haplotype sharers, divided by the number of haplotype sharers.
- percentage of aligned sequences in the first round of an algorithm. Out of all detected IBD regions in the first round, what proportion of them were aligned into one of the gametes.

A.6.1. Reference sharing in ORCADES study

The evaluation of the algorithms was possible thanks to parent offspring pairs genotyped in the ORCADES study. There are 58 individuals with both parents genotyped, and at 80% of their heterozygous loci they could be phased. There are 160 with at least one parent genotype and they could be phased at 70% of heterozygous loci.

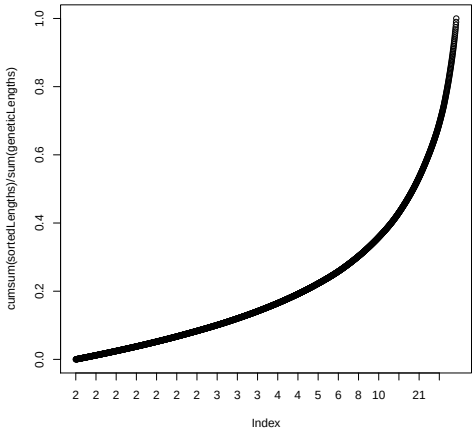
To obtain the reference IBD information, we extracted IBD regions between the 58 reliably phased reference individuals. We required alleles with identical alleles in a region larger than 2 cM and containing at least 100 SNPs. The length of IBD regions between the reference individuals is shown in Figure 5a. The density of IBD sharing across the genome is shown in Figure 5c.

We also counted IBS sequences shorter than 2 cM. Contrary to our estimates about the expected lengths of IBD since re-settling of the island around 50 generations ago, there are many such segments. The lengths of IBS segments shorter than 2 cM are presented in Figure 5b.

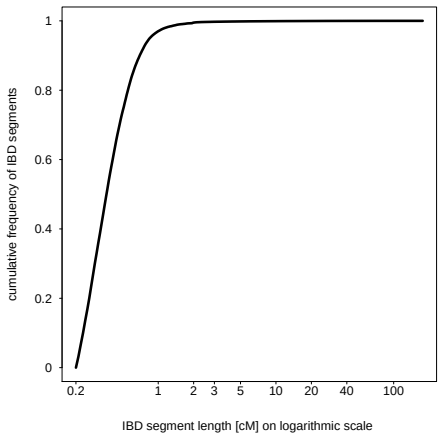
A.6.2. IBD threshold in Stage I

In Stage I of ANCHAP we would like to phase the individuals widely with as few phasing errors as possible. Genotypes would be widely phased if many haplotype sharers are widely detected. There would be few phasing errors if there is no falsely detected IBD sharing. Therefore the sensitivity and false discovery rate of IBD detection, evaluated on the reference phased individuals, are meaningful metrics which will reflect the quality of phasing.

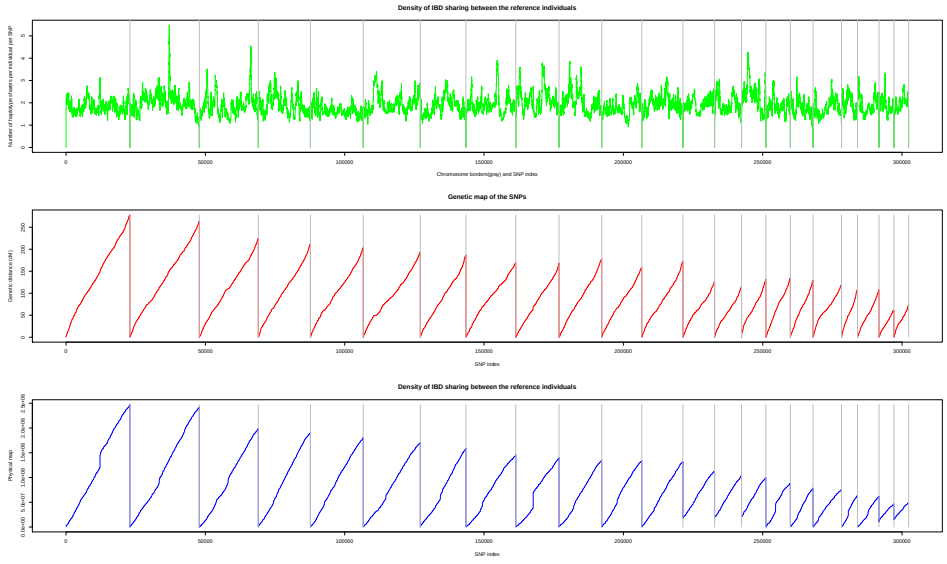
On the other hand, when there is more than one haplotype sharers, and some falsely detected IBD region, the alignment stage of Anchap may eliminate the falsely detected



(a) How long are the regions of reference sharing? Horizontal axis - region length in cM on log-scale. Vertical axis - cumulative proportion of of IBD segments in the IBD regions.



(b) Length of IBD segments in reference sharing. Most of the segments are much shorter than 2 cM.



(c) How is the reference sharing distributed across the chromosome? Is it influenced by fluctuations in the genetic or physical map? Top: density of IBD between the 58 reference individuals in ORCADES. Middle: genetic map. Bottom: physical map.

Fig. 5.—: Reference data for our comparison: sequences shared between the individuals which can be reliably phased

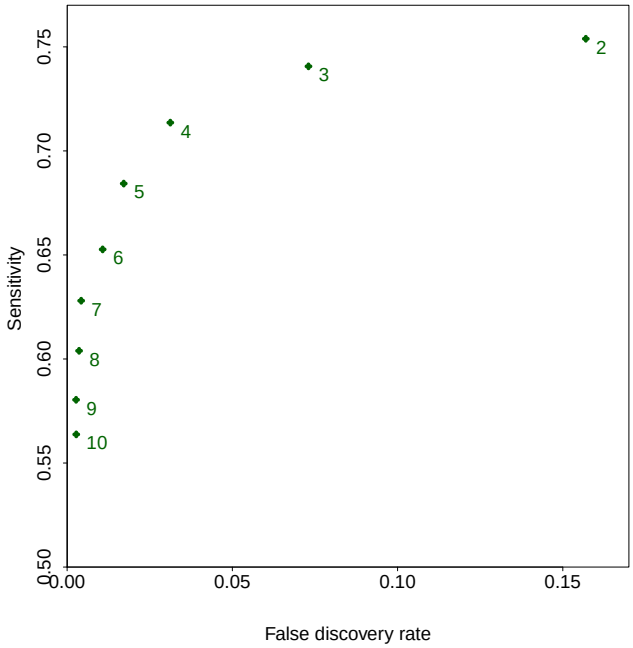


Fig. 6.—: Sensitivity and false positive rate of IBD regions as recovered by Stage I of ANCHAP with different IBD thresholds.

sharing.

The plot of sensitivity and false discovery rates for different IBD thresholds in the first round are shown in Figure 6.

A.6.3. IBD region margins

At each border of the putative sharing regions we trimmed 100 markers. In the experiments with the reference data, after trimming 100 markers at each side 94% of detected sharing regions will not contain any spurious sharing at the borders.

A.6.4. Stage II - alignment parameters

In Stage II of ANCHAP haplotype sharers are split into two groups, and regions of falsely assumed sharing may be discarded. The algorithm starts with the longest and therefore most certain IBD regions, reconstructs a draft of the phase, and then matches the remaining sharers against the preliminary phased genotypes. Errors may occur in the preliminarily reconstructed haplotypes, and therefore few inconsistencies between the draft of the haplotypes and the aligned sequences may be allowed.

There are two parameters necessary for this part of the algorithm. The overlap threshold specifies the minimal number of markers of overlap between the draft of phase of an individual and the new IBD region. The matching threshold specifies how many alleles may be mismatching between the draft of the phase and a genotype of the putative IBD sharer.

Right values of parameters will result in a good split of haplotype sharers into two groups and consequently to low phasing error, and a good proportion of the genotypes will be phased. A good proportion of the putative IBD sequences would be aligned. The genotypes of IBD sharers who are all classified as sharing the same haplotype should also be consistent between each other. There should be no opposing homozygotes between such genotypes, and therefore the inconsistency ratio should be low.

In Table 1 we evaluate the impact of different values of the overlap threshold and the matching threshold. For each pair of values, we evaluate the percentage of the putative IBD regions successfully aligned, and the inconsistency ratio.

overlap threshold	matching threshold	percentage of haplotype sharers aligned	inconsistent homozygotes among haplotype sharers, normalised
5	0	0.07	8.90E-06
5	0.01	0.67	5.11E-04
5	0.02	0.74	1.00E-03
5	0.05	0.82	2.12E-03
5	0.1	0.86	3.45E-03
5	0.2	0.91	5.55E-03
5	0.5	0.50	3.74E-03
10	0	0.07	8.92E-06
10	0.01	0.67	5.14E-04
10	0.02	0.73	1.00E-03
10	0.05	0.86	2.01E-03
10	0.1	0.86	3.46E-03
10	0.2	0.90	5.56E-03
10	0.5	0.50	3.74E-03
20	0	0.07	8.92E-06
20	0.01	0.66	5.19E-04
20	0.02	0.73	1.01E-03
20	0.05	0.81	2.15E-03
20	0.1	0.86	3.49E-03
20	0.2	0.90	5.59E-03
20	0.5	0.49	3.77E-03
50	0	0.07	8.92E-06
50	0.01	0.64	5.35E-04
50	0.02	0.71	1.05E-03
50	0.05	0.78	2.21E-03
50	0.1	0.83	3.55E-03
50	0.2	0.87	5.67E-03

Table 1:: Experiments with parameters for Stage II of ANCHAP. According to these parameters it is decided whether two diplotype segments are aligned, ie. whether they share the same gamete. Marked in gray is the value of the parameter used.

A.6.5. Stage III parameters

In Stage III we look for haplotypes matching continuously in regions which are at least 2cM long. In addition we require that both of the compared haplotypes are phased - another parameter specifies a minimum number of markers phased in both of the haplotypes - by default it is set to 100. In Table 2 we show the accuracy of IBD detection when this threshold is varied. For sensitivity the threshold has a negligible impact, while false discovery rate increases significantly when the threshold is set to less than 100.

lengththresh	sensitivity	false discovery rate
10	0.84	0.031
20	0.84	0.031
50	0.84	0.027
100	0.84	0.016
200	0.81	0.011

Table 2:: Experiments with values of the parameter of the second round scan - lengththresh. This parameter specifies how many markers in the region of putative IBD need to be phased. Marked in gray is the value of the parameter used.

A.7. Challenges to haplotype alignment

The quality of haplotype reconstruction, as measured by the switch error, is influenced by accuracy of sharing detection and of the algorithm that splits haplotype sharers into two groups. ANCHAP's greedy algorithm first chooses the longest shared sequences, as they carry most information about the haplotype, and tries to align the haplotypes from the remaining shared regions. When the proband's haplotype is not phased in the region, the assignment to a haplotype is arbitrary, and these arbitrary decisions may not be propagated between genetic regions.

A.8. Tuning SLRP

Table 3 shows experiments with empirical and default parameter values for SLRP. We compared the default values with values obtained empirically. The expected IBD length in centimorgans was computed from the IBD regions between the reference individuals in ORCADES, after they were phased. The expected IBS but not IBD was calculated from IBS segments between the reference individuals, longer than 20 markers. Because we defined IBD as matching of haplotypes within a region longer than 2 cM, out of the output of SLRP we filtered out the results shorter than this threshold.

SLRP setting	ExpectedIBS (cM)	ExpectedIBD (cM)	sensitivity	false discovery rate
default	1	10	0.76	0.0076
empirical	0.42	9.17	0.77	0.0106

Table 3:: Tuning SLRP. Only counting the IBD regions longer than 2cM. Sharing between the 58 Orkney individuals was evaluated. Data from chromosome 2. Marked in gray is the value of the parameter used.

A.9. Tuning fastIBD

In Table **??** we show experiments with varying the scale parameter in fastIBD. We filtered out regions shorter than 2 cM, in accordance with our definition of IBD.

A.10. Comparison of Anchap, SLRP and fastIBD

In the article we evaluated IBD regions as inferred by different regions against the IBD segments between the reference individuals. Here we additionally show density of IBD across the genome (Figure 7) and comparison of lengths of detected IBD segments (Figure 9).

fastIBD setting	scale	sensitivity <2cM pruned	false discovery rate <2cM pruned	sensitivity	false discovery rate
minimum advised	1	0.270	0.000	0.271	0.002
	2	0.631	0.010	0.635	0.023
	2.5	0.744	0.018	0.750	0.036
	2.6	0.767	0.018	0.774	0.037
	2.7	0.783	0.019	0.790	0.043
	2.8	0.802	0.021	0.808	0.046
	2.9	0.805	0.024	0.812	0.051
	3	0.825	0.024	0.832	0.056
	3.1	0.832	0.027	0.839	0.066
	3.2	0.837	0.028	0.844	0.070
	3.3	0.845	0.030	0.853	0.073
	3.4	0.849	0.032	0.855	0.079
	3.5	0.857	0.036	0.865	0.088
maximum advised	4	0.870	0.045	0.879	0.118
merge 10 runs	3	0.868	0.044	0.873	0.106

Table 4:: Tuning fastIBD. Sharing between the 58 Orkney individuals was evaluated. Data from chromosome 2. Marked in gray is the value of the parameter used.

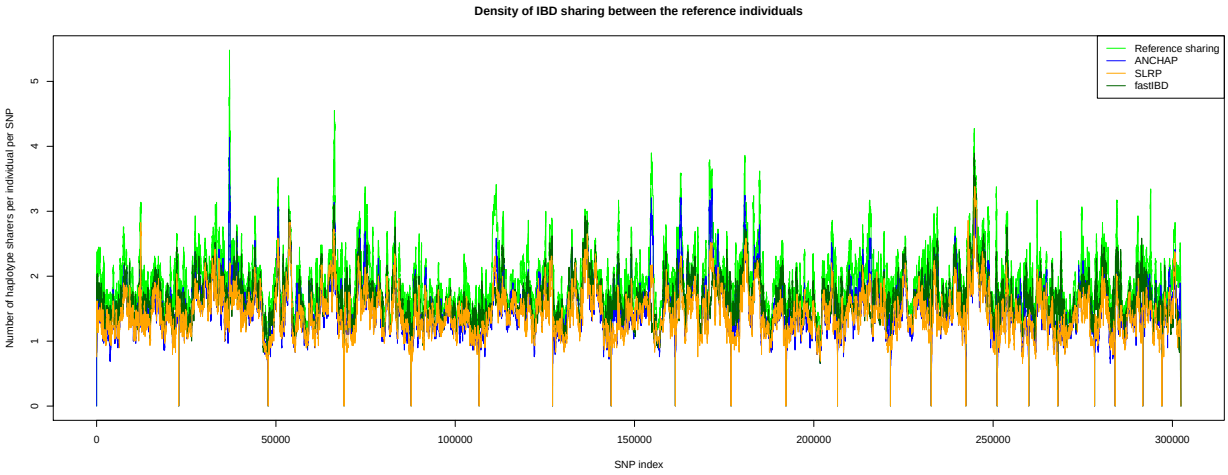


Fig. 7.—: Genome-wide view of haplotype sharing as recovered by the compared methods. SLRP and fastIBD are more conservative in IBD detection, and have less apparent IBD peaks.

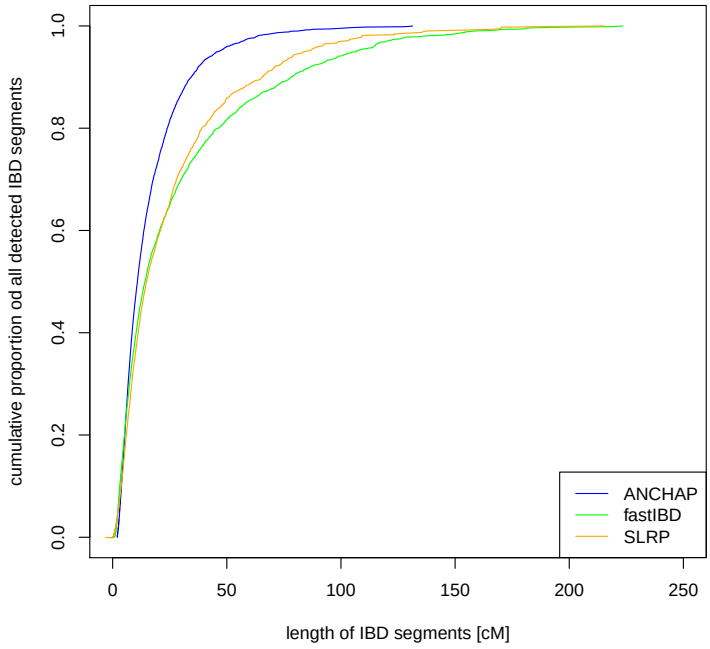


Fig. 8.—: Lengths of detected IBD segments [cM]. IBD regions detected by ANCHAP are generally shorter, as the method does not account for switch errors in phasing after the first round.

B. Evaluation of the selection procedure

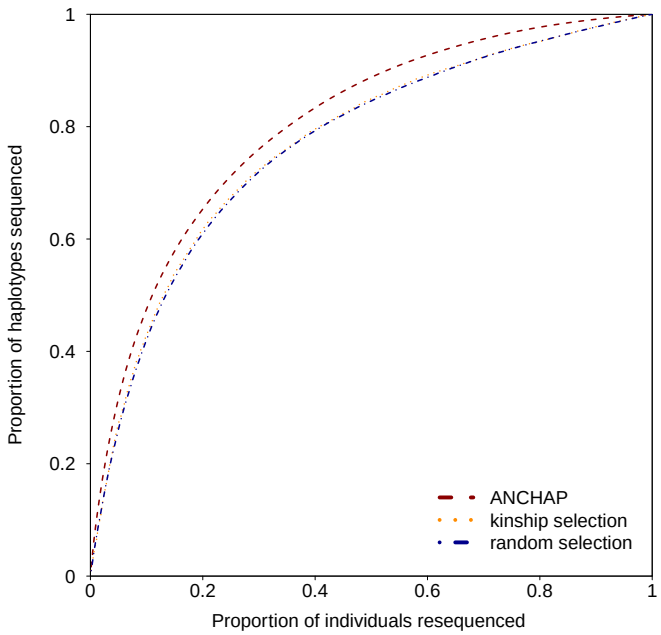


Fig. 9.—: Evaluation of ANCHAP's selection procedure for choosing subjects for resequencing studies. We estimated the IBD imputation potential when the samples are chosen randomly or based on kinship. Individuals were chosen randomly 10 times, and the results were averaged.

REFERENCES

I. H. Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–861, Oct 2007. doi: 10.1038/nature06258. URL http://dx.doi.org/10.1038/nature06258.

K. Palin, H. Campbell, A. F. Wright, J. F. Wilson, and R. Durbin. Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genet Epidemiol*, Oct 2011. doi: 10.1002/gepi.20635. URL http://dx.doi.org/10.1002/gepi.20635.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–575, Sep 2007. doi: 10.1086/519795. URL http://dx.doi.org/10.1086/519795.

This manuscript was prepared with the AAS LaTeX macros v5.0.